Traitement automatique des termes composés : segmentation, traduction et variation

Elizaveta Loginova Clouet

Thèse de Doctorat

Elizaveta LOGINOVA CLOUET

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de l'Université de Nantes Angers Le Mans

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique
Spécialité : Traitement Automatique du Langage Naturel
Unité de recherche : Laboratoire d'informatique de Nantes-Atlantique (LINA)

Soutenue le 19 décembre 2014

Traitement automatique des termes composés : segmentation, traduction et variation

JURY

Présidente :	**M^me^ Natalie KÜBLER**, Professeur des universités, Université Paris Diderot
Rapporteurs :	**M. Nabil HATHOUT**, Directeur de recherche, CNRS et Université de Toulouse
	M^me^ Natalie KÜBLER, Professeur des universités, Université Paris Diderot
Examinateur :	**M^me^ Natalia GRABAR**, Chargée de recherche, CNRS et Université de Lille 1&3
Directrice de thèse :	**M^me^ Béatrice DAILLE**, Professeur des universités, Université de Nantes

Remerciements

Je remercie avant tout Béatrice Daille, ma directrice de recherche, de m'avoir accueillie en thèse et de m'avoir offert cette belle opportunité d'acquérir des connaissances en informatique et terminologie tout en travaillant sur un sujet relevant du traitement des langues. Je suis très reconnaissante envers elle pour ses conseils précieux, sa disponibilité, sa confiance et son encadrement expérimenté.

Je remercie également Natalie Kübler et Nabil Hathout d'avoir accepté d'être rapporteurs pour cette thèse, et d'avoir porté un regard extérieur sur mon travail avec des remarques stimulantes. Je tiens à souligner que Natalie Kübler a aussi participé à mon comité de suivi de thèse et que dans ce cadre, nous avions eu des échanges intéressants et fructueux au cours de ma thèse. Je remercie Natalia Grabar d'avoir accepté d'être examinatrice (il me tenait à coeur qu'il y ait une personne d'origine slave dans ce jury !) et d'avoir apporté des remarques approfondies pendant la soutenance. Je remercie chaleureusement le deuxième membre de mon comité de suivi de thèse, Evelyne Jacquey, pour sa lecture attentive et pédagogique de mon manuscrit en fin de première et deuxième années de thèse.

Cette thèse a débuté dans le cadre du projet européen TTC, et c'était une expérience formidable pour moi de participer à un projet dans lequel collaboraient plusieurs équipes scientifiques et entreprises. Je remercie de tout coeur tous nos partenaires dans le projet, et particulièrement Helena Blancafort, Ulrich Heid, Serge Sharoff et Tatiana Gornostay : il était très enrichissant pour moi de travailler à leur côtés. Je remercie également les personnes de l'équipe TALN ayant participé à ce projet, notamment Emmanuel Morin et Jérôme Rocheteau.

Je suis reconnaissante à toute l'équipe TALN qui m'a accueillie. C'est une équipe de gens extrêmement sympathiques, ouverts, passionnés par ce qu'ils font et d'une curiosité impressionnante. Je remercie, entre autres Emmanuel P., Florian, Nicolas, Chantal... J'ai aussi une pensée toute particulière pour Sacha : avec son aide je suis entrée dans cette équipe et dans ce laboratoire.

Le LINA est un laboratoire multinational et multiculturel, c'est un espace extrêmement accueillant. Je remercie les personnes administratives et techniques du laboratoire qui aident tous les chercheurs au quotidien. J'ai eu la chance de travailler dans ces conditions très agréables, entourée par d'autres doctorants (pour une partie d'entre eux, aujourd'hui docteurs) qui sont devenus mes amis : Rima, Nagham, Ophélie, Audrey, Amir, Prajol, Mathieu V., Benjamin, Thomas V., Adrien, Firas et bien d'autres.

Je remercie Estelle Delpech, qui a également soutenu sa thèse dans cette équipe, parce que son travail m'a servi d'exemple, et de plus, j'ai utilisé dans mes expériences les tables de correspondance de morphèmes qu'elle avait construites. Je remercie de nouveau Rima Harastani, cette fois pour notre travail en collaboration, et aussi pour ses listes d'éléments néoclassiques que j'ai utilisées. Mes remerciements vont également à Franziska Duckstein qui a annoté, avec beaucoup d'efficacité, les données en allemand qui m'ont servi de référence dans mes expériences avec cette langue.

Enfin, je remercie mon mari Samuel de sa relecture de ma thèse, et encore plus de son soutien tout au long de ce travail. Je n'en dis pas plus, il le sait.

Table des matières

Liste des tableaux

Table des figures

Notation

Langues	
DE	Allemand
EN	Anglais
ES	Espagnol
FR	Français
JP	Japonais
IT	Italien
RU	Russe
Catégories grammaticales	
N	Nom
ADJ	Adjectif
V	Verbe
VPART	Participe du verbe
ADV	Adverbe
PREP	Préposition
CONJ	Conjonction
NP	Groupe nominal
Annotation morphologique	
PREF	Préfixe
SUF	Suffixe
DES	Désinence
SG	Singulier
PL	Pluriel
GEN	Génitif
Abréviations	
TAL	Traitement Automatique des Langues
TAS	Traduction Automatique Statistique
RTL	Reference Term List (Liste terminologique de référence)
SWT	Single Word Term (terme monolexical)
MWT	Multi Word Term (terme polylexical)
CM	Composé morphologique
CS	Composé syntaxique
TS	Terme source (terme de la langue source)
TC	Terme cible (terme de la langue cible)

– C'est une belle chose, la destruction des mots. Naturellement, c'est dans les verbes et les adjectifs qu'il y a le plus de déchets, mais il y a des centaines de noms dont on peut aussi se débarrasser. Pas seulement les synonymes, il y a aussi les antonymes. Après tout, quelle raison d'exister y a-t-il pour un mot qui n'est que le contraire d'un autre ? Les mots portent en eux-mêmes leur contraire. Prenez « bon », par exemple. Si vous avez un mot comme « bon » quelle nécessité y a-t-il à avoir un mot comme « mauvais » ? « Inbon » fera tout aussi bien, mieux même, parce qu'il est l'opposé exact de bon, ce que n'est pas l'autre mot. Et si l'on désire un mot plus fort que « bon », quel sens y a-t-il à avoir toute une chaîne de mots vagues et inutiles comme « excellent », « splendide » et tout le reste ? « Plusbon » englobe le sens de tous ces mots, et, si l'on veut un mot encore plus fort, il y a « doubleplusbon ». Naturellement, nous employons déjà ces formes, mais dans la version définitive du novlangue, il n'y aura plus rien d'autre. En résumé, la notion complète du bon et du mauvais sera couverte par six mots seulement, en réalité un seul mot. Voyez-vous, Winston, l'originalité de cela ?

George Orwell, « 1984 »

1

Introduction

Un illustre scientifique français du 18e siècle, Pierre Bouguer, menait des recherches à la fois en mathématiques, physique, hydrographie, géographie et architecture navale. Son contemporain, un scientifique russe Mikhaïl Lomonossov, est reconnu pour ses travaux en chimie, physique, astronomie, géographie, géologie et pédagogie. De nos jours, il serait très difficile d'envisager une telle polyvalence, car les connaissances ont considérablement évolué, les sciences se sont diversifiées et spécialisées. Par exemple, la physique englobe de nombreuses branches comme la mécanique, l'optique, l'acoustique, l'électricité, le magnétisme, la physique quantique et bien d'autres. De plus, les technologies jouent un rôle majeur dans la société contemporaine. De nouveaux domaines de spécialité apparaissent en permanence, et les domaines existants évoluent eux aussi sous cette influence. On assiste en conséquence à la prolifération des termes dans l'espace informationnel, dans la littérature spécialisée ou vulgarisée, dans les média et sur le web. Cette croissance est si rapide que les organismes de systématisation et de normalisation de la terminologie parviennent à peine à suivre le rythme.

Face à cette situation, la *terminographie* (l'activité de recensement et de gestion des termes) s'est emparée des dernières avancées en informatique (on parle même parfois de *terminotique* ou encore de *terminologie computationnelle*). Actuellement, des outils informatiques interviennent aux diverses étapes du traitement des termes et de la construction des ressources terminologiques. Quant à la terminologie computationnelle (« *computational terminology* », nommée sur le modèle de « *computational linguistics* »), il s'agit d'un domaine de recherche qui se situe au carrefour de la terminographie et du traitement automatique des langues (TAL), et dont le but est de faire avancer les techniques du TAL qui pourront servir à des fins terminographiques. Ce travail de thèse s'inscrit dans ce domaine.

La communication autour d'un sujet spécialisé, pour être précise et efficace, nécessite de la part du récepteur la connaissance de la terminologie utilisée par l'émetteur. La communication est d'autant plus compliquée quand les intervenants n'ont pas de langue natale commune. Pour rendre la communication spécialisée la plus efficiente possible, des ressources terminologiques mono- et multilingues sont créées, souvent au format électronique. Ces ressources sont destinées aux traducteurs, spécialistes du domaine, étudiants, novices, etc. Des applications pour la rédaction ou la traduction, l'indexation automatique et d'autres applications du TAL font également appel à de telles ressources.

La première étape de la création d'une ressource terminologique est la sélection de l'ensemble des termes à décrire, que nous appelons ici *lexique terminologique*. La construction d'un lexique terminologique à partir d'une collection de textes relatifs à un domaine donné, *corpus spécialisé*, constitue un axe de

recherches en terminologie computationnelle : *l'extraction terminologique*. Cette thèse est effectuée dans le cadre du projet européen TTC « Terminology Extraction, Translation Tools and Comparable Corpora » qui vise le développement d'outils d'extraction terminologique monolingue et bilingue, destinés à améliorer les systèmes de traduction automatique statistique, les systèmes d'aide à la traduction et ceux de gestion de terminologie.

Problématique et objectifs

En terminologie, on distingue les termes simples des termes complexes (du point de vue sémantique). En TAL, on manipule des unités graphiques. La forme et le sens ne coïncident pas toujours : parfois un terme sémantiquement complexe est réalisé par un seul mot graphique (ce qu'on appelle *terme composé*). Les différentes réalisations des termes complexes selon les langues ont déjà été pointées par les travaux en traduction automatique statistique (Koehn et Knight, 2003; Fritzinger et Fraser, 2010). Dans cette thèse, nous souhaitons démontrer que les limites des notions « terme simple » et « terme complexe » sont relatives et dépendent de la langue, et que cette distinction ne suffit pas pour classer les termes si on veut appliquer les mêmes techniques de traitement aux différentes langues.

Notre objectif est de montrer que les termes composés constituent une catégorie importante des termes dans de nombreuses langues et nécessitent un traitement particulier pour les reconnaître et les segmenter en parties constituantes sémantiquement autonomes afin de pouvoir leur appliquer les techniques du TAL. La composition étant un des principaux procédés de formation des mots dans la langue générale, le traitement des mots composés dépasse les limites de la terminologie compositionnelle. Dans le même temps, ce phénomène est particulièrement productif dans les langues de spécialité. Dans cette thèse, nous nous focalisons sur le traitement des termes composés pour la construction de lexiques bilingues.

Pour atteindre notre objectif, nous commençons par l'évaluation des résultats d'une extraction automatique de termes. Pour cette évaluation, nous construisons de manière semi-manuelle des listes terminologiques de référence. Nous construisons des listes pour deux langues, le français et l'anglais, ainsi que des listes bilingues anglais-français et français-russe pour deux domaines de spécialité : *l'énergie éolienne* et *les technologies mobiles*. La comparaison entre les difficultés rencontrées pendant cette construction semi-manuelle et les erreurs commises par un extracteur de termes permettra de souligner les problèmes spécifiques à l'extraction automatique, et notamment les problèmes liés au traitement des termes composés.

Nous proposons ensuite une méthode d'identification et de segmentation automatique des termes composés. Jusqu'au début des années 2000, la segmentation était effectuée sur la base de règles implémentant des théories linguistiques. Cette approche est très précise, mais spécifique à une langue donnée. Depuis ces dernières années, les corpus de textes sont sollicités pour la tâche de segmentation. Les méthodes exploitant les corpus sont indépendantes de la langue, mais moins précises. Nous tentons de concilier les deux avantages dans une méthode fondée sur l'utilisation des corpus et capable d'intégrer des règles linguistiques. De plus, nous souhaitons qu'elle soit adaptable aux domaines de spécialité. Cette méthode sera évaluée sur quatre langues - anglais, allemand, français et russe - et deux domaines - *l'énergie éolienne* et *le cancer du sein*.

Notre deuxième objectif est de démontrer en quoi la segmentation des termes composés peut être bénéfique pour la construction de lexiques terminologiques. Nous prenons l'exemple de deux tâches : l'alignement (la traduction) des termes composés et le regroupement des variantes terminologiques. Pour ces expériences, un cadre unifié est proposé : quatre langues (les mêmes que celles retenues pour les expériences de segmentation), deux domaines de spécialité (*l'énergie éolienne* et *le domaine médical*) et les résultats produits par notre système de segmentation.

Un lexique bilingue établit des liens entre les termes de la langue source et leurs équivalents de la langue cible. La segmentation des termes composés d'une langue source ayant une composition productive permet de mettre en correspondance les composants identifiés avec les mots graphiques constituant des termes multi-mots (nous les appellerons *polylexicaux*) dans une langue cible dans laquelle la composition est moins productive. Pour démontrer cela, nous appliquons une *approche compositionnelle* de la traduction des unités lexicales complexes.

Un lexique monolingue peut être enrichi par des *variantes* de termes. Un terme apparaît dans les textes spécialisés sous des formes variées, et pourtant l'utilisateur (ou l'application finale) a besoin de le reconnaître. La segmentation des termes composés peut servir pour l'identification de leurs variantes polylexicales dans les textes. Nous menons des expériences en mettant en place une technique basique du regroupement des variantes qui n'exploite pas de patrons pré-établis de la variation, ce qui la rend indépendante de la langue et permet d'extraire des variantes ayant des structures différentes qui ne sont pas définies d'avance.

Structure de la thèse

Après cette introduction, nous passerons en revue quelques notions de base de la terminologie computationnelle et du TAL nécessaires pour décrire nos travaux, dans le chapitre 2. Le chapitre 3 de cette thèse sera consacré à l'extraction terminologique monolingue et bilingue. Dans le chapitre 4 nous discuterons de l'évaluation de l'extraction terminologique, particulièrement celle utilisant une liste de termes de référence. Nous décrirons notre expérience de construction de telles listes et nous commenterons les résultats d'une évaluation effectuée avec les listes construites auparavant.

Dans le chapitre 5 nous nous concentrerons sur le phénomène de la composition dite « morphologique » et nous examinerons des méthodes de l'état de l'art de la segmentation des composés. Le chapitre 6 présentera notre méthode de segmentation, qui sera ensuite évaluée et comparée avec trois autres méthodes de l'état de l'art.

Les expériences de traduction compositionnelle des termes composés identifiés et segmentés seront rapportées dans le chapitre 7. La variation des termes composés sera examinée dans le chapitre 8. Le chapitre 9 conclura cette thèse.

En Annexes on trouvera la description des ressources terminologiques et des outils cités dans ce travail (annexe A), ainsi que la description des ressources utilisées dans les expériences (dictionnaires, corpus, lexiques) (annexe B). Les paramètres retenus pour la segmentation sont résumés dans l'annexe C. L'annexe D présente des extraits des résultats de la segmentation pour un domaine avec la segmentation de référence. Les patrons de la variation observés sont décrits dans l'annexe E.

2

Notions de base

Nous commençons par introduire quelques notions-clés de la terminologie computationnelle et du traitement automatique des langues. Ce sont les notions de terme (section 2.1), de langue de spécialité et de spécificité du terme (section 2.2), de ressource terminologique (section 2.3) et de corpus de textes (section 2.4). Nous aborderons également certains aspects de la formation de nouveaux termes (section 2.5) et la variation terminologique (section 2.6).

2.1 Terme

La notion centrale de la terminologie est le **terme**. Néanmoins, une définition universelle du terme n'existe pas. Reprenons quelques définitions proposées dans la littérature :

1. « nom définissable à l'intérieur d'un système cohérent, énumératif (nomenclature) ou structuré (taxonomie) et correspondant sans ambiguïté à une notion ou concept » (Rey-Debove, 1979) ;
2. « unité signifiante constituée d'un mot ou de plusieurs mots et qui désigne une notion de façon univoque à l'intérieur d'un domaine » (Boutin-Quesnel et al., 1985) ;
3. « mot considéré dans sa valeur de désignation, en particulier dans un vocabulaire spécialisé » (Dubois, 1980) ;
4. « unité lexicale dont le sens est envisagé par rapport à un domaine de spécialité ». (L'Homme, 2004).

Ces définitions ont en commun le fait qu'un terme n'existe pas « tout seul », ainsi il est toujours défini par rapport à son domaine de spécialité, c'est-à-dire « un domaine de la connaissance humaine, souvent associé à une activité socio-professionnelle » (L'Homme, 2004, p. 22). Ainsi, Cabré (1998, p. 94) définit la terminologie comme « l'ensemble des mots spécialisés d'une discipline donnée (ou d'un domaine d'activités) ». Otman (1996, p. 15) écrit également : « le terme, pour être circonscrit, n'a pas besoin d'un contexte (comme le mot) mais d'un domaine d'appartenance ». Le **domaine de spécialité** est définitoire du terme.

Approche conceptuelle (onomasiologique) vs. contextuelle (sémasiologique). Dans le même temps, les définitions citées relèvent deux visions différentes de la terminologie, parfois appelées « conceptuelle » ou « onomasiologique » et « contextuelle » ou « sémasiologique ». D'après l'**approche onomasiologique**

(du grec *onoma* 'nom'), les termes sont considérés comme des étiquettes attribuées aux concepts, unités mentales existant a priori. Le but de l'analyse terminologique est alors de révéler la structure conceptuelle du domaine : on part donc ici des concepts pour aller vers les termes. Dans les années 1930, quand la terminologie en tant que science est apparue grâce aux travaux de Wüster (1931), cette vision était dominante. Cependant, elle a été critiquée dans des recherches plus récentes (Otman, 1996; Cabré, 1998), surtout du fait de l'intérêt croissant pour la terminologie multilingue. En linguistique, l'intuition suivante a déjà été évoquée avec l'hypothèse de Sapir-Whorf (Sapir, 1929; Whorf, 1956) : des langues différentes découpent le monde différemment, et la langue que l'on parle influence notre conceptualisation de la réalité. Un concept « naturel » pour un locuteur natif d'une langue n'a pas forcément d'équivalent exact et univoque dans la représentation du monde du locuteur d'une autre langue. L'exemple classique est la variété des mots pour désigner des formes différentes de neige dans les langues des peuples esquimaux (Whorf, 1956). La terminologie (dans le sens de « l'ensemble des termes d'un domaine ») reflète le même relativisme, malgré l'effort considérable de la communauté scientifique de la normaliser. Par exemple, le *mot*, une des notions fondamentales de la linguistique, n'a pas la même interprétation quand il s'agit de langues isolantes, agglutinantes ou fusionnelles. Un exemple du domaine de l'énergie éolienne : EN *windmill* est l'équivalent en FR de *moulin à vent*, c'est-à-dire le dispositif utilisé essentiellement pour moudre les grains et non pour produire de l'électricité. Dans l'usage, ce terme est toutefois souvent utilisé également pour identifier l'outil qui s'appelle en français *aérogénérateur* ou EN *wind generator*, i.e. un dispositif de production d'électricité. En français, le terme *moulin à vent* est rarement utilisé pour parler des *aérogénérateurs*, par contre le terme plus général *éolienne* peut être utilisé [1].

L'**approche sémasiologique** (du grec *sêma* 'signe') est apparue en terminologie comme une réponse à la critique de l'approche onomasiologique et suite au rapprochement de la terminologie avec la linguistique (Condamines, 2005). Cette approche, contrairement à l'approche onomasiologique, part des formes linguistiques pour remonter aux concepts. Les termes sont tout d'abord des unités lexicales et leur sens ne peut être établi que dans un contexte et par rapport aux autres unités lexicales de la langue (cf. les définitions (3) et (4) du terme) ; l'étude des termes doit donc se faire sur la base des emplois réels des termes, i.e. des textes spécialisés. Dans la même optique, la « terminologie textuelle » se développe (Bourigault et Slodzian, 1999). Comme son nom l'indique, cette approche part des textes pour décrire les termes. Pour les représentants de ce courant, une ressource terminologique doit être construite à partir de l'utilisation réelle des termes et pour un objectif concret.

De nos jours, les deux visions se complètent et cohabitent dans les pratiques terminologiques. Notre travail, s'inscrivant dans le cadre du TAL, se situe dans l'approche contextuelle (sémasiologique) et dans la tradition d'analyser des unités lexicales issues de textes spécialisés. Cette approche est plus adaptée au cadre multilingue car elle ne présuppose pas l'existence de concepts universels.

Univocité et monoréférentialité du terme. Un terme peut avoir plusieurs définitions selon le domaine. Pour illustrer ceci, le terme *lacet* dans le Grand Dictionnaire Terminologique [2] est attaché aux domaines suivants : 1) *chaussure* (« sorte de cordon plat ou rond, fait de soie, de coton, etc., habituellement terminé par des ferrets ») ; 2) *industrie de la confection* (« cordelette coulissante servant à resserrer diverses pièces d'habillement ») ; 3) *loisir* (« piège de rétention comportant un fil métallique, un ressort, une détente à palette et qui est destiné à retenir un animal à fourrure par la patte ») ; 4) *marine* (« mouvement non désiré d'un bateau autour de son axe vertical, causant l'oscillation de l'avant et de l'arrière du bâtiment de gauche à droite ») ; 5) *cybernétique* (« dans un plan analogue à celui qui est formé par le plat d'une main humaine, rotation de l'organe terminal effecteur autour d'un axe perpendiculaire à ce plan ») ; 6) *aéronautique* (« mouvement d'un avion autour d'un axe vertical passant par le centre de gravité »), etc.

Dans la tradition onomasiologique, l'ambiguïté des termes à l'intérieur d'un domaine est déclarée impossible. Les définitions du terme (1) et (2) citées ci-dessus soulignent l'univocité et la monoréférentialité

1. TERMIUM Plus, http://www.btb.termiumplus.gc.ca/, date de consultation 02.10.2014.
2. http://www.gdt.oqlf.gouv.qc.ca, date de consultation : 18.07.12

en tant que propriétés du terme (un terme correspond à un concept, et réciproquement). Ces principes théoriques, par conséquent, excluent la possibilité de la polysémie et de la synonymie entre les termes. Pour les partisans de l'approche sémasiologique, ces principes sont contestés par les attestations issues des textes. La polysémie et la synonymie entre les termes sont possibles même au sein d'un domaine et dans la même langue, par exemple, dans l'usage des courants concurrents ou à cause de la métaphorisation ou de la vulgarisation du terme (Otman, 1996). Ainsi, le terme linguistique français *catégorie grammaticale* est synonymique au terme *partie du discours*. Dans le domaine de l'énergie renouvelable, l'expression *tour éolienne* peut être utilisée comme synonyme de *mât éolien* [3]. Le verbe EN *hire*, qui est un terme fréquent du domaine du commerce, est polysémique et peut signifier 'louer' (*to hire a car* - 'louer une voiture') ou 'embaucher' (*to hire staff* - 'embaucher du personnel'). L'existence de la polysémie et de la synonymie entre les termes dans les textes spécialisés contribue aussi à expliquer pourquoi nous adhérons à l'approche contextuelle (sémasiologique).

Termes simples, complexes, monolexicaux et polylexicaux. Puisque dans l'approche contextuelle le terme est une unité lexicale, nous nous intéressons à la forme des termes. Du point de vue de la forme, on distingue les **termes simples** et les **termes complexes**. Un terme simple est un « terme constitué d'un seul radical avec ou sans affixes » (AFNOR, 1990) : *rotor, éolienne*. Un terme complexe est un « terme constitué de deux ou plusieurs radicaux auxquels peuvent s'ajouter d'autres éléments » (ibid.) : *centrale éolienne, pale de rotor, insulino-dépendant, électrothérapie*.

Dans la littérature anglophone, deux notions légèrement différentes sont d'usage : **« *single-word term (SWT)* »** et **« *multiword term (MWT)* »**. La première consiste en un seul mot graphique, tandis que la deuxième en comporte plusieurs (les mots graphiques sont délimités à l'écrit par des espaces ou d'autres séparateurs de mots). Parfois ces notions (SWT vs. MWT) sont utilisées comme équivalentes à celles de termes simples et complexes. Pour nous, les termes tels que *électrothérapie* et *insulino-dépendant* sont complexes, mais comportent chacun un seul mot graphique. Les deux distinctions sont utiles en traitement automatique des termes : la première (termes simples vs. complexes) traduit la complexité/simplicité du sens, la deuxième - celle de la forme. Le traitement d'un terme est différent selon qu'il contient une ou plusieurs mots graphiques. Pour garder cette nuance, nous utilisons les dénominations **terme monolexical** (*rotor, électrothérapie, insulino-dépendant*) et **terme polylexical** (*centrale éolienne, pale de rotor*) comme correspondant respectivement aux SWT et MWT.

Cette convention permet de décrire l'existence des termes à la fois complexes et monolexicaux (*électrothérapie, insulino-dépendant*), qui nécessitent une segmentation avant de les traiter. Les termes complexes polylexicaux, quant à eux, posent le problème de leur identification en tant qu'entités dans les textes. Prenons comme exemple la phrase « il existe plusieurs méthodes pour assurer la continuité de service d'un système d'information ». La *continuité de service d'un système d'information* est-elle un terme complexe, ou est-ce plutôt *service d'un système d'information* ? ou bien encore *continuité de service* et *système d'information* séparément ? Un humain dira que c'est plutôt cette dernière proposition qui est correcte, mais cela représente un défi pour les systèmes automatiques. Pour caractériser à quel point les parties d'un syntagme sont sémantiquement liées et forment une unité, Kageura et Umino (1996) utilisent la notion de *degré d'unité* (« *unithood* »). Les termes polylexicaux ont donc un haute degré d'unité. Des techniques statistiques, linguistiques et mixtes ont été proposées pour mesurer ce degré d'unité, cf. section 3.1.2.

3. TERMIUM Plus, http://www.btb.termiumplus.gc.ca/, date de consultation 29.09.2014.

2.2 Langue de spécialité et spécificité du terme

Les termes d'un domaine font partie de la langue de spécialité. La langue de spécialité ou autrement dit la langue spécialisée (terme anglais LSP : « *language for special purpose* ») est une autre notion très importante et en même temps largement discutée en terminologie. Sager et al. (1980, p. 21) la définit comme étant les « moyens de communication linguistique requis pour véhiculer de l'information spécialisée parmi les spécialistes d'une même matière ». Lerat (1995, p. 20) critique cette définition, parce que la langue spécialisée peut être utilisée non seulement par les professionnels d'un domaine, mais aussi par leurs clients, le public, etc. Alors il propose la définition suivante de la langue spécialisée : « une langue naturelle considérée en tant que vecteur de connaissances spécialisées ».

La langue de spécialité est souvent définie par opposition à la langue générale (LGP, « *language for general purposes* »). Bowker et Pearson (2002) écrivent : « la langue générale est la langue qu'on utilise tous les jours pour parler des choses ordinaires dans la variété des situations communes, la langue de spécialité est utilisée pour parler des domaines spécialisés des connaissances » [4]. Cependant, la frontière entre LSP et LGP n'est pas parfaitement claire. Cabré (1998, p. 224) indique la gradualité de la spécialisation : « On peut considérer que les « langues » de la physique, de la chimie, de la biologie, de la géologie, des mathématiques, des statistiques, de la linguistique, de l'anthropologie, de l'histoire, de l'architecture, de l'économie théorique, etc., présentent un très haut degré de spécialisation. Il s'agit clairement de langues de spécialité. D'autres « langues », comme celles de la banque, de la Bourse, du droit et de l'économie appliquée constituent un terrain intermédiaire entre les langues plus spécialisées et celles plus générales. Celles de la restauration, de la coiffure, de la ferronnerie, des sports, enfin, présentent un degré bien moindre de spécialisation, et se trouvent de ce fait à la frontière de la langue commune ». Dans le même temps, la LSP doit à la LGP une grande partie du vocabulaire et de la syntaxe, et de ce fait elle est parfois définie comme une sous-partie de la langue générale (Kocourek, 1991; Cabré, 1998), et non comme son contraire.

Un des critères du texte spécialisé est son sujet (scientifique, technique ou professionnel). Dans un cadre communicatif, Cabré (1998, p. 125) ajoute deux autres critères à celui du *sujet* du texte : *les utilisateurs* et *les situations de communication*. Elle accentue le côté pragmatique : l'émetteur est spécialiste, le récepteur peut être spécialiste ou non initié ; la communication est normalement de type formel et « régie par des critères professionnels ou scientifiques ».

Également dans une perspective communicative, une autre notion qui se croise avec la LSP a été proposée : *la communauté de discours* (Swales, 1990). Derrière cette notion il y a l'idée que les gens réunis par des objectifs communs utilisent dans leur discours des moyens linguistiques et rhétoriques particuliers. Swales (1990) distingue six caractéristiques significatives identifiant une communauté de discours :

- des objectifs communs admis publiquement (ex. pour les membres d'une association de jardinage : améliorer leurs techniques de jardinage) ;
- des mécanismes de communication entre les membres (les gardiens de phare dans des endroits éloignés et ne communiquant pas ne forment pas une communauté de discours) ;
- des mécanismes de participation (liste de diffusion, forum, etc.) servent principalement à avoir un retour d'information. Si une personne paie son inscription annuelle à une association sans pour autant participer à ses activités, la personne n'appartient pas à cette communauté de discours ;
- l'utilisation des genres spécifiques ;
- la terminologie spécifique ;
- le niveau élevé d'expertise (il y a des experts avec leurs connaissances).

L'auteur donne un exemple intéressant qui n'est ni scientifique, ni technique, ni professionnel, mais qui répond aux six critères identifiés d'une communauté de discours : « Hong Kong Study Circle », l'association des amateurs de timbres postaux de Hong Kong. Il s'agit d'une autre façon de parler d'une langue spécialisée.

4. « LGP is the language we use every day to talk about ordinary things in a variety of common situations. In contrast, LSP is used to discuss specialized fields of knowledge » (Bowker et Pearson, 2002, p. 25).

En fait, il est plus correct de parler non pas *d'une* mais *de langues de spécialité*, car il y a de nombreux domaines de savoir, chacun avec sa propre terminologie (Bowker et Pearson, 2002). Néanmoins, toutes les LSP partagent des caractéristiques communes, notamment (Cabré, 1998) : la fonction informative comme leur fonction prioritaire, la concision, la précision, le caractère international.

Une autre caractéristique traditionnellement retenue des langues de spécialité est leur impersonnalité. Cependant, les recherches récentes montrent que les LSP sont beaucoup moins impersonnelles qu'on ne le pensait (Hunston et Thompson, 2000). Malgré leur forme impersonnelle générale, les langues de spécialité peuvent comporter un style personnalisé qui apparaît, entre autres, dans le dialogue implicite entre l'auteur et le lecteur. Il y a également certaines thématiques où la prise de position est inévitable (e.g. le nucléaire).

Il est important de souligner la différence entre une langue de spécialité et la terminologie d'un domaine. La LSP n'est pas une simple combinaison de vocabulaire général et de terminologie, elle contient également des règles de syntaxe particulières (par exemple, l'emploi très fréquent des constructions passives), des traits stylistiques et des moyens différents d'organiser l'information par rapport à la LGP (Bowker et Pearson, 2002). Les LSP ont tendance à utiliser des structures morphologiques complexes (ex. *hyperéosinophilie*), abréviations, sigles, symboles spécialisés (π), emprunts d'autres langues (anglais, latin), à susciter la domination des formes nominales et la nominalisation des verbes, etc. (Cabré, 1998). Les termes sont alors des éléments du vocabulaire des LSP.

Tous les mots qui apparaissent dans les textes spécialisés ne sont pas des termes. Selon Jackendoff (1983), un terme peut être défini par une suite de conditions nécessaires et suffisantes, tandis que dans la langue générale un concept est défini par rapport à un prototype. Par exemple, un *oiseau* dans la langue générale est défini par son degré de ressemblance avec l'oiseau-prototype (le moineau), et dans la zoologie il est identifié par l'existence de plumes. Ceci est un exemple de l'approche onomasiologique : le terme d'une langue est considéré comme une constante qui peut être définie indépendamment de son emploi selon le contexte. Dans la démarche sémasiologique, Kageura et Umino (1996) introduisent la notion de *degré de spécificité* (« *termhood* »), i.e. « le degré par lequel une unité linguistique est liée aux concepts spécifiques d'un domaine [5] ». Par exemple, l'expression *continuité de service* est un terme du domaine des technologies mobiles, mais l'unité lexicale *groupe de services* n'en est pas un, car elle est utilisée dans les textes de domaines très variés et dans la vie quotidienne. Quant à l'expression *mise en service*, son statut reste contestable, parce qu'elle n'est pas spécifique à un domaine en particulier, mais largement utilisée dans les situations liées aux technologies et aux systèmes. Nous allons suivre cette démarche et considérer comme termes toutes les unités lexicales avec un fort degré de spécificité au sein d'un domaine.

Dans un contexte applicatif, la question se pose donc : comment mesurer le degré de spécificité d'une unité lexicale ? Les expériences montrent que même les experts évaluant les candidats-termes ont un degré d'accord généralement assez bas : ainsi, Vivaldi et Rodríguez (2007) décrivent leur expérience dans laquelle l'accord entre trois experts participant à l'évaluation a été constaté pour seulement 37 % des candidats termes. Cette tâche est donc d'autant plus difficile à automatiser. Certaines techniques statistiques sont néanmoins proposées pour essayer de résoudre ce problème (cf. section 3.1.1).

2.3 Ressource terminologique

Des diverses ressources répertorient et décrivent des termes. Les ressources terminologiques principales sont des dictionnaires spécialisés, de format papier ou électronique, qui réunissent les termes d'un domaine de spécialité, et des banques terminologiques, de format électronique, qui regroupent les termes appartenant aux domaines variés (L'Homme, 2004).

Les données dans les ressources terminologiques sont organisées d'une façon différente des dictionnaires généraux. Parmi les modèles utilisés selon les ressources, on constate aussi des variations. Pour faci-

5. « Termhood refers to the degree that a linguistic unit is related to domain-specific concepts » (Kageura et Umino, 1996)

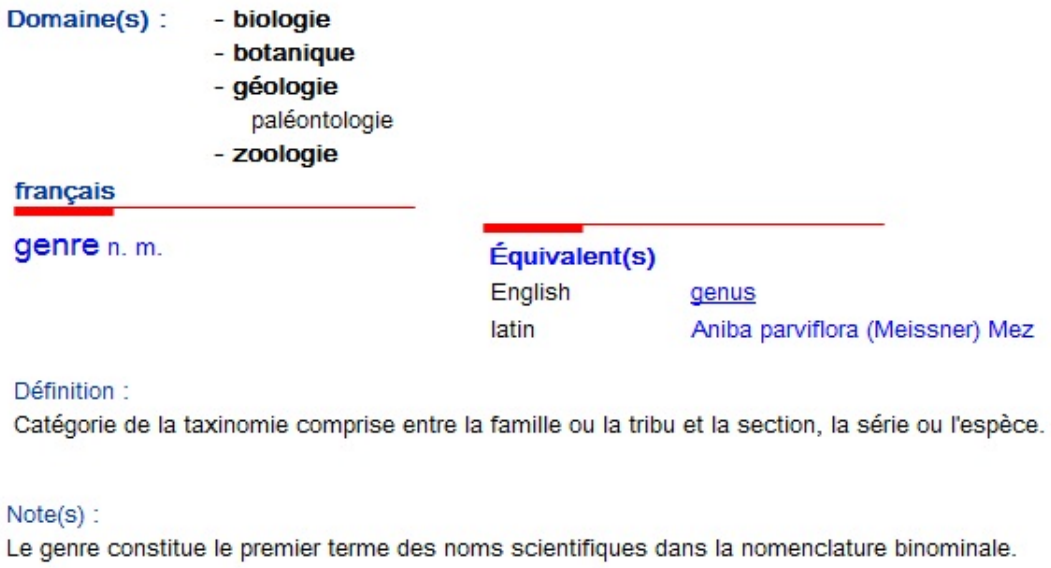

FIGURE 2.1 – Fiche terminologique *genre* de Grand Dictionnaire Terminologique

liter la navigation d'une ressource à une autre en vue d'un travail terminologique, un organisme de normalisation a été créé - « ISO/TC 37 Terminologie et autres ressources langagières et ressources de contenu » [6]. Cet organisme édite des normes de description des données terminologiques.

Une entrée d'une ressource terminologique (une fiche terminologique) contient d'habitude le terme en question, l'annotation linguistique (catégorie grammaticale, genre pour les noms, transitivité pour les verbes, etc. en fonction de la langue), le domaine, la définition, une note (commentaire), les contextes d'utilisation. Les illustrations, la prononciation, les marques d'usage (*local, jargon, soutenu, néologisme*, etc.) et d'autres informations complémentaires sont des informations potentiellement présentes en fonction des objectifs de la ressource. Une fiche peut aussi inclure des références à d'autres termes liés par des relations conceptuelles (« espèce - genre », « tout-partie », etc.) ou lexico-sémantiques (hyponymie, méronymie, synonymie, antonymie, etc.). Elle peut également mentionner des liens vers des équivalents dans d'autres langues si la source est multilingue. On se reportera à l'exemple d'une fiche *genre* du domaine biologique extraite du Grand Dictionnaire Terminologique [7], cf. figure 2.1. Les banques terminologiques et les dictionnaires électroniques renvoient à la source à partir de laquelle le terme a été extrait. De plus, ils contiennent souvent des méta-données (la référence d'une source bibliographique, la date de saisie d'une fiche, etc.).

Un des axes de recherches en terminologie computationnelle est l'acquisition de la terminologie à partir des textes spécialisés. Les termes extraits forment un lexique terminologique. Le lexique terminologique n'est pas encore une ressource terminologique telle qu'elle a été définie, mais il peut être utilisé pour la construction d'un dictionnaire ou d'une banque terminologique, ou il peut être intégré dans une application de gestion de la terminologie ou de traduction assistée par ordinateur.

6. www.iso.org/iso/fr/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=48104
7. www.granddictionnaire.com/btml/fra/r_motclef/index800_1.asp, date de consultation : 15.03.12.

2.4 Corpus

L'utilisation des corpus est devenue quasi systématique en TAL ainsi qu'en pratiques lexicographiques parce qu'elle permet de travailler avec le matériel concret et « vivant » de la langue, des données dites « authentiques ». De la même manière, en terminologie computationnelle, l'utilisation des corpus de langue spécialisée permet de se baser sur des exemples réels d'emploi des termes.

Le corpus est défini en TAL comme « une collection de données langagières sélectionnées et ordonnées d'après des critères linguistiques explicites afin d'être utilisée comme échantillon représentatif d'une langue » [8] (Sinclair, 1996). Les données langagières réunies dans les corpus sont soit des textes écrits, soit des enregistrements audio ou vidéo (les corpus oraux et multimédia). Dans les pratiques terminologiques il est d'usage d'exploiter les corpus de textes écrits.

Actuellement les corpus existent au format électronique. Ils sont généralement associés à des interfaces plus ou moins élaborées (parfois appelées *concordanciers*) permettant de lancer des recherches dans les données. En plus de la recherche dite « plein texte », une recherche plus sophistiquée est possible grâce à l'annotation linguistique (morphologique, syntaxique, etc.) et métalinguistique (auteur, année, etc.).

On distingue les **corpus généraux** des **corpus spécialisés**. Un corpus général est conçu pour être le plus représentatif possible par rapport à une langue naturelle dans son ensemble (exemples : BNC [9], Corpus National Tchèque [10], Corpus National Russe [11]). Un corpus spécialisé est représentatif d'un sous-ensemble d'une langue (d'un genre, d'une période, d'un auteur, d'un phénomène particulier, etc.).

Si les données rassemblées dans un corpus appartiennent à une langue, on parle de **corpus monolingue**. Il existe également des **corpus bilingues** ou **multilingues**. Parmi les corpus multilingues, on distingue les corpus parallèles et comparables. Les **corpus parallèles** sont constitués de textes dans une langue source associés à leur traduction dans une (ou des) langue(s) cible(s). Les **corpus comparables** sont des « collections de textes de langues différentes qui ne sont pas des traductions » [12] (Bowker et Pearson, 2002, p. 93). Pour mesurer le degré de comparabilité des corpus, des critères stylistiques (les textes partageant le même sujet, genre, période, média, registre, etc.), ainsi que des critères quantitatifs (fréquences de mots) sont utilisés (Déjean et Gaussier, 2002).

Dans les pratiques terminologiques, les corpus spécialisés du domaine étudié sont utiles pour pouvoir observer les termes du domaine et leurs contextes d'utilisation. Les corpus généraux sont toutefois aussi utilisés en tant que corpus de référence pour révéler quels phénomènes sont caractéristiques de la langue générale et quels phénomènes sont, au contraire, spécifiques au domaine donné. Les corpus monolingues et multilingues sont sollicités selon que la tâche à effectuer porte sur une ou plusieurs langues. Actuellement les corpus sont exploités par les terminologues pour illustrer l'usage des termes, pour formuler les définitions et pour construire la terminologie d'un domaine. Il existe de plus en plus d'outils d'aide à la rédaction des terminologies qui, à partir des corpus de textes spécialisés, proposent à l'utilisateur des candidats-termes et regroupent des termes potentiellement liés. L'utilisation des corpus en terminologie n'est, néanmoins, qu'une pratique assez récente et liée au paradigme contextuel (Condamines, 2005).

8. « A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language » (Sinclair, 1996)
9. British National Corpus http://www.natcorp.ox.ac.uk
10. Český národní korpus http://ucnk.ff.cuni.cz
11. Corpus National Russe http://ruscorpora.ru
12. « Sets of texts in different languages, that are not translations of each other » (Bowker et Pearson, 2002, p. 93).

2.5 Formation de nouveaux termes

En terminologie, la production de nouveaux mots - la néologie - est extrêmement présente. Comme l'écrit Cabré (1998, p. 252) : « la néologie, conçue comme activité de création de nouvelles dénominations, s'impose dans les domaines de spécialité où l'apparition constante de nouvelles notions exige une créativité lexicale permanente ». En terminologie, la néologie est désignée par le terme de *néonymie* (Rondeau, 1983).

Cabré (1998) distingue deux grands types de néologismes : référentiels (qui désignent une nouvelle notion) et expressifs (qui introduisent une nouvelle forme dans les buts d'expression, pour souligner une nuance ou un aspect de la notion existante). Elle résume les critères proposés pour déterminer le statut néologique d'une unité lexicale :

— unité apparue récemment ;
— unité ne figurant pas dans les dictionnaires (ou dans un corpus admis comme corpus de référence) ;
— unité présentant des signes d'instabilité formelle ou sémantique ;
— unité nouvelle dans la perception des usagers.

Les néonymes, contrairement aux néologismes de la langue générale, doivent aussi répondre aux critères des termes, donc, selon le critère communément admis, ils doivent appartenir à un domaine de spécialité.

Pecman (2012) met l'accent sur l'instabilité en tant que caractéristique clé des néologismes, y compris des néonymes, qui sont largement sujets à la synonymie. L'ignorance de ce fait (selon l'approche conceptuelle, les termes sont caractérisés par la stabilité et évitent la synonymie) vient du manque d'attention portée au rôle expressif des néonymes. En étudiant la néologie expressive sur le matériel du discours scientifique, l'auteur révèle une fonction rhétorique des néonymes qui consiste en une promotion, probablement inconsciente, du côté novateur de la pensée scientifique.

La construction de termes, en tant qu'unités lexicales, s'effectue par les mêmes voies que celle des autres unités de la langue (par la dérivation, la composition, etc.), mais chaque procédé n'a pas la même productivité. La **réduction ou branchygraphie** (sigles, acronymes, abréviation) et les **emprunts à d'autres langues** sont des procédés minoritaires dans la langue générale, mais dans la terminologie, soumise à l'apparition permanente de néologismes, ils sont très fréquents : *igbt-diode*, RU ip-адрес 'adresse ip', etc.

La dérivation est un mécanisme très courant dans la formation de termes simples. Elle se définit en linguistique comme la procédure qui « produit un nouveau mot à partir d'un seul mot préexistant en le modifiant » (Lehmann et Martin-Berthet, 2008, p.147). La modification peut porter sur la forme du mot ou/et sur sa catégorie grammaticale. Ainsi, on distingue :

1. L'affixation, qui procède par ajout d'un affixe (*histoire - préhistoire, neurone - neuronale*), parfois accompagné par une altération du radical (*cancer - cancéreux*). La catégorie grammaticale du dérivé peut être la même que celle de la base ($histoire_N$ - $préhistoire_N$) ou différente d'elle ($neurone_N$ - $neuronale_{ADJ}$, $cancer_N$ - $cancéreux_{ADJ}$). Les *adjectifs dénominaux*, i.e. dérivés de nom, sont extrêmement productifs dans les textes spécialisés ;
2. La dérivation parasynthétique, quand un mot est formé par un ajout simultané du préfixe et du suffixe (ou désinence) au radical : *câble* → *encablure* (les formes **encâble*, **cablure*, **encabler* n'existent pas en français) (Lehmann et Martin-Berthet, 2008) ;
3. La recatégorisation ou autrement dit la conversion qui consiste à changer la catégorie grammaticale sans ajout d'affixe (Neveu, 2004) : *la phonétique*$_N$ - *phonétique*$_{ADJ}$;
4. La dérivation régressive (Neveu, 2004) qui consiste à obtenir un nouveau mot à partir d'un autre par suppression d'un affixe (suffixe ou désinence) : $embranchement_N \rightarrow embrancher_V$, $aider_V \rightarrow aide_N$. Ce type peut être vu comme un cas particulier de recatégorisation.

Les mots dérivés d'une même racine se regroupent dans un paradigme dérivationnel : *allergie - allergique - allergène* ; *fraise* (dans le sens 'outil d'usinage') - *fraisage - fraiseuse - fraisé*, etc (L'Homme, 2004).

La formation de termes complexes se fait par **la composition** de deux ou plusieurs éléments lexicaux sémantiquement autonomes, « mots à *sens plein* » (Lehmann et Martin-Berthet, 2008), représentés par des formes dites populaires (natives) - EN *toolbar* 'barre d'outils', - ou savantes (néoclassiques) - *homolatéral*. La composition dans un sens restreint, c'est-à-dire qui inclut seulement les unités dont les éléments sont concaténés, est plus ou moins productive selon les langues, mais lorsqu'elle est possible dans une langue, elle est toujours particulièrement présente dans les domaines de spécialité. La composition dans un sens élargi, i.e. incluant les unités dont les éléments sont séparés par des espaces (*vitesse de rotation*, EN *rotational speed*, RU ветроэнергетический ресурс 'ressource d'énergie éolienne'), est également un des traits caractéristiques des terminologies (Cabré, 1998), i.e. elle est aussi plus fréquente dans les langues de spécialité que dans la langue générale. Dans ce travail, nous aborderons principalement le traitement des termes composés dans le sens restreint.

La formation de termes se poursuit en permanence dans les langues. De nouveaux domaines apparaissent, les domaines existants évoluent avec leur terminologie. Le flux de nouvelles données et par conséquent de nouveaux termes engendre la coexistence de variantes terminologiques.

2.6 Variation

Malgré le principe de stabilité de la forme du terme établi par l'approche conceptuelle, les études récentes montrent que dans les textes réels les termes ont tendance à varier. La variation est un phénomène non négligeable car un grand nombre d'occurrences des termes dans les textes constituent en fait des variantes. Afin de donner un ordre de grandeur, dans un corpus scientifique anglais, les variantes représentent un tiers de toutes les occurrences des termes d'après l'étude de Jacquemin (1999).

La variation terminologique est un phénomène complexe. Sur le plan synchronique, la variation peut être vue comme une des relations entre les termes, mais une relation qui existe dans le discours et non dans le système de la langue. Cependant les variantes souvent employées se lexicalisent avec le temps et entrent dans les dictionnaires (Depierre, 2007). Ainsi, sur le plan diachronique, la variation se croise avec la néonymie. Les recherches sur le sujet sont relativement récentes, et la perception de ce phénomène varie selon les auteurs. On distingue la variation conceptuelle de la variation linguistique (dénominative).

La variation conceptuelle consiste en des changements progressifs qui apparaissent au sein d'un concept. On la constate avec des degrés divers d'équivalence entre les significations d'une unité lexicale et entre les significations de ces variantes (Kostina, 2011). Par exemple, dans les deux phrases « en mettant la table, assurez-vous que les couteaux sont propres » et « l'intrus m'a menacé avec un couteau » (les exemples sont tirés du (Kostina, 2011)), le mot *couteau* a des significations extrêmement proches mais pas absolument identiques. La variation conceptuelle découle des processus polysémiques que les termes subissent dans le discours.

La variation linguistique (dénominative) consiste, avant tout, en des changements de dénominations. Freixa (2006) la définit comme un « phénomène dans lequel le même concept a des dénominations différentes » [13]. L'auteur précise que par dénominations différentes elle entend les formes lexicalisées, donc relativement stables et répandues dans le domaine. Cette définition est en fait assez proche de celle de la synonymie, plus précisément, de la quasi-synonymie. Une définition similaire se trouve dans le travail de Depierre (2007) : l'auteur considère comme des variantes de termes toutes les formes graphiques similaires, ainsi que toutes les formes ayant une signification similaire. Par conséquent, les formes ayant une signification différente sont traitées comme des termes différents.

Nous allons suivre une conception plus élargie de la variation dénominative (Daille et al., 1996; Jacquemin, 1999). D'après cette vision, « la variante d'un terme est un énoncé qui est sémantiquement et

13. « Denominative variation can be defined as the phenomenon in which one and the same concept has different denominations » (Freixa, 2006, p. 51).

conceptuellement lié au terme d'origine »[14] (Daille et al., 1996), mais qui n'est pas forcément son synonyme : « une variante peut être synonymique avec le terme de référence, ou montrer une certaine distance sémantique par rapport à ce terme (tout en gardant la référence au même concept ou à un concept proche) » (Daille, 2005).

Plusieurs typologies de variation peuvent être proposées en fonction de l'application finale (la recherche d'information, l'extraction de terminologie, l'indexation semi-automatique, systèmes de question-réponse, etc.). Les types principaux selon Daille (2005) sont :

1. Variation graphique : la modification de la forme graphique (tiret, alternance "s"/"z" en EN, etc.) *kilowattheure - kilowatt-heure* ;
2. Variation flexionnelle : la modification de désinence *conservation de produit - conservation de produits* ;
3. Variation syntaxique : le changement de la structure syntaxique de l'unité terminologique par l'insertion d'un élément ou la permutation des composants *protéine végétale - protéine d'origine végétale*, EN *hand function - function of the hand* ;
4. Variation morphosyntaxique : le changement de la structure syntaxique et de la forme morphologique *acidité du sang - acidité sanguine* ;
5. Variation paradigmatique : la substitution d'un ou de plusieurs composants de l'unité terminologique par son synonyme en préservant la structure syntaxique *épuisement du combustible - appauvrissement du combustible*.

Les variantes syntaxiques et morphosyntaxiques ne sont pas toujours en relation de synonymie avec leur terme de base, elles peuvent entretenir des relations d'hyperonymie, d'antonymie, etc. Nous citons quelques exemples : *huile essentielle - huile essentielle de sapin* (variante syntaxique), *brunissement enzymatique - brunissement non enzymatique* (variante syntaxique), *sucre de betterave - sucrerie de betterave* (variante morphosyntaxique). De telles variantes dépassent le cadre des définitions de la variation données par Freixa (2006) et Depierre (2007). Elles sont cependant utiles pour certaines applications terminologiques telles que l'établissement des relations entre les termes, la construction des ontologies, etc. Elles montrent aussi la productivité du terme. Les variantes syntaxiques et morphosyntaxiques ont été étudiées par Ibekwe-SanJuan (1998), Jacquemin et Tzoukermann (1999), Daille (2003), Grabar et Zweigenbaum (2004) et d'autres chercheurs.

Freixa (2006) a fait une étude des causes de la variation dénominative en terminologie. Elle relève six types majeurs de causes :

1. Causes préliminaires telles que la redondance linguistique et le caractère arbitraire du signe linguistique ;
2. Causes dialectiques (facteurs géographiques, chronologiques et sociaux) ;
3. Causes fonctionnelles (l'adaptation de l'émetteur du message au niveau de langage et de spécialisation du récepteur) ;
4. Causes discursives (le besoin d'éviter la répétition, la tendance vers « l'économie » de l'expression linguistique, et l'autre tendance vers la créativité) ;
5. Causes inter-linguistiques telles que la coexistence d'un terme natif avec un emprunt d'une autre langue ;
6. Causes cognitives (imprécision conceptuelle, facteur « idéologique », différence dans la conceptualisation des phénomènes).

14. « A variant of a term is an utterance which is semantically and conceptually related with an original term » (Daille et al., 1996, p. 201)

La variation terminologique a été étudiée principalement du point de vue monolingue. Des études existent pour les langues suivantes (Daille, 2005) : le français, l'anglais, l'allemand, l'italien, l'espagnol et le catalan, le japonais, le malgache, le polonais. Dans une perspective multilingue les études ne sont pas nombreuses et ne couvrent généralement que des couples de langues incluant l'anglais.

Dans cette thèse, nous aborderons plus précisément un type de variation dénominative terminologique, celle entre les termes complexes monolexicaux et leurs variantes polylexicales. Nous étudierons la productivité de ce type de variantes et leurs structures morphosyntaxiques dans les langues que nous traitons.

2.7 Bilan

Ce tour d'horizon de quelques notions de la terminologie et du traitement automatique des langues nous a permis d'argumenter notre positionnement dans le paradigme contextuel de la terminologie. Nous pouvons à présent nous intéresser aux aspects techniques du traitement des termes, qui sont utiles dans le cadre de la construction des ressources terminologiques.

La construction d'une ressource terminologique comprend la sélection des unités lexicales ayant un statut terminologique dans un domaine donné (ou plusieurs domaines donnés), la recherche des définitions et des contextes pour illustrer l'usage des termes, l'établissement des relations plus ou moins fines entre les termes, etc. Le travail de cette thèse s'inscrit dans le cadre de l'extraction terminologique, i.e. la sélection des termes de manière automatique. Cette tâche inclut aussi l'établissement de quelques liens entre les termes extraits, notamment de lien d'équivalence entre les termes en différentes langues ou de lien entre les termes et leurs variantes au niveau monolingue. Ces aspects seront abordés par la suite.

I

Contexte applicatif : l'extraction terminologique et son évaluation

La construction d'une ressource terminologique commence par la sélection des termes à décrire. Dans l'optique contextuelle, des termes d'un domaine sont extraits pour ce but à partir d'un corpus de textes spécialisés relevant de ce domaine. L'acquisition de la terminologie d'un domaine forme tout un axe de recherche en terminologie computationnelle et en traitement automatique des langues. Elle peut se faire pour une langue donnée (extraction monolingue) ou pour plusieurs langues (extraction bilingue et multilingue). Dans ce dernier cas, ce processus comprend l'établissement d'équivalence (quand cela est possible) entre les termes sélectionnés des différentes langues.

De nombreuses recherches ont été effectuées dans cet axe, et des outils d'extraction automatique ont été mis en place. Plusieurs stratégies sont appliquées pour évaluer et comparer les différentes méthodes d'extraction. Nous nous intéressons plus précisément à une des stratégies d'évaluation, celle utilisant une liste terminologique de référence. Une liste de référence des termes d'un domaine peut être construite manuellement à partir d'un ensemble de textes spécialisés. Sa construction se rapproche de l'extraction automatique des termes, sauf qu'elle est effectuée par un humain.

Dans cette partie de la thèse, nous abordons les deux sujets : les aspects théoriques et techniques de l'extraction automatique des termes, ainsi que la problématique de construction des listes de référence dans le cadre de l'évaluation des extracteurs. La partie pratique de ce travail consiste à construire de telles listes pour deux langues (FR, EN) et deux domaines de spécialité (l'énergie éolienne, les technologies mobiles). Enfin, nous rapportons les résultats de l'évaluation d'un extracteur terminologique (TermSuite) effectuée en utilisant les listes obtenues, et nous apportons quelques remarques sur les erreurs typiques commises par l'extracteur automatique, et d'une manière plus générale, sur l'évaluation à l'aide de liste de référence.

3

Extraction terminologique

Dans ce chapitre, nous résumons les techniques principales d'extraction automatique des termes d'un domaine donné à partir d'un corpus de textes. Dans un cadre monolingue (section 3.1), nous examinons séparément l'extraction des termes monolexicaux (3.1.1) et polylexicaux (3.1.2) à partir de corpus de textes, puis l'extraction des termes (mono- et polylexicaux) en utilisant des ressources sémantiques (3.1.3). Dans un cadre bilingue (section 3.2), nous examinons l'extraction terminologique à partir de corpus parallèles (3.2.1) et comparables (3.2.2).

3.1 Extraction terminologique monolingue

L'extraction des termes monolingues consiste à repérer dans une langue donnée les termes relevant d'un domaine de spécialité, à partir d'un corpus de textes ou/et en utilisant des ressources sémantiques.

3.1.1 Extraction des termes monolexicaux

Un des indices de base pour mesurer le lien entre une unité lexicale et un domaine est sa fréquence dans un corpus spécialisé de ce domaine. La fréquence absolue d'une unité lexicale dans un corpus correspond au nombre d'occurrences de cette unité dans le corpus. La fréquence relative d'une unité lexicale X est égale à sa fréquence absolue divisée par le nombre total de mots dans le corpus.

Avec une approximation élevée on peut dire que plus une unité lexicale est fréquente dans un corpus spécialisé, plus elle est caractéristique du domaine en question (plus son statut terminologique est sûr). Le corpus doit nécessairement être lemmatisé au préalable et « nettoyé » des mots grammaticaux (articles, prépositions, auxiliaires, etc.). Ainsi, pour repérer les candidats termes d'un domaine, il faut indexer les mots du corpus spécialisé, supprimer les mots usuels, et classer les mots restants par fréquence décroissante : avec une probabilité élevée, les termes les plus pertinents se trouveront en tête de cette liste.

Cependant, les fréquences des mots varient beaucoup dans la langue générale, et cela a un impact sur les langues de spécialité. Pour retenir les termes, il ne faut pas seulement sélectionner les mots les plus fréquents, mais aussi tenir compte de leur spécificité par rapport au domaine. La spécificité d'une unité lexicale X peut être calculée de plusieurs manières, par exemple en divisant sa fréquence relative dans

le corpus spécialisé $Freq_{LSP}$ par sa fréquence relative dans un corpus général utilisé comme corpus de référence $Freq_{LGP}$ (« *weirdness ratio* » (Ahmad et al., 1992)) :

$$\text{Spécificité(X)} = \frac{Freq_{LSP}(X)}{Freq_{LGP}(X)} \tag{3.1}$$

Pour extraire les candidats termes d'un corpus spécialisé, on peut donc filtrer le lexique du corpus par la spécificité des candidats termes.

Les filtrages par fréquence et par spécificité sont des fonctions courantes pour les concordanciers et les extracteurs de termes : WordSmith Tools [1], AntConc [2], IMS Open Corpus Workbench [3], etc.

Ces techniques sont devenues traditionnelles dans les pratiques terminologiques car elles sont relativement simples et efficaces. Cependant, elles ont aussi leurs limites et leurs restrictions. Premièrement, elles peuvent être appliquées seulement à condition de disposer d'un corpus spécialisé d'une taille importante, parce que la corrélation entre la spécificité et la fréquence devient stable à partir d'un volume de données important. Deuxièmement, les corpus doivent être lemmatisés. De plus, l'homonymie et la polysémie langagières peuvent modifier considérablement les statistiques. Par exemple, *l'entrée* est un terme important en informatique, pourtant le calcul de la fréquence ainsi que de la spécificité pour ce mot seront biaisés à cause de (1) l'homonymie entre le nom *entrée* et le participe passé du verbe *entrer*, et (2) la polysémie entre le mot *entrée* dans son emploi spécialisé ('entrées d'un outil') et son emploi général ('entrée de la maison'). Enfin, on ne peut jamais garantir si un candidat est un terme monolexical ou seulement une partie d'un terme polylexical. Ainsi, les mots *axe* et *vertical* dans le corpus de l'énergie éolienne ne sont pas des termes autonomes, mais des éléments d'un terme polylexical *axe vertical*.

3.1.2 Extraction des termes polylexicaux

Il est admis que la majeure partie des termes dans les textes sont des termes polylexicaux (i.e. constitués de plusieurs unités graphiques). L'extraction des termes polylexicaux est toutefois une tâche plus fine que l'extraction des termes monolexicaux, car le degré spécificité d'un syntagme est difficile à mesurer, et en plus, il faut mesurer le degré d'unité (« *unithood* » (Kageura et Umino, 1996)) pour bien identifier les frontières d'un terme polylexical. Cette tâche peut être accomplie en s'appuyant également sur le calcul des fréquences et en exploitant des heuristiques linguistiques.

Pour estimer le degré d'unité des syntagmes et identifier les syntagmes avec un haut degré d'unité, les mesures d'association sont exploitées. Nous citons ici deux mesures d'association couramment utilisées en linguistique de corpus et en terminologie computationnelle, *l'information mutuelle* et *log-likelihood*.

Ces mesures s'appuient sur deux valeurs de probabilité de co-occurrence de deux mots : la probabilité obtenue, i.e. attestée dans le corpus, et la probabilité espérée, i.e. attendue dans le cas où il n'y a pas de corrélation entre deux mots.

L'*information mutuelle* entre X et Y est calculée ainsi (Church et Hanks, 1990) :

$$IM(X,Y) = \log_2 \frac{P(X,Y)}{P(X)P(Y)} \tag{3.2}$$

où $P(X,Y)$ renvoie à la probabilité obtenue de co-occurrence de deux mots X et Y, et $P(X)$ renvoie à la probabilité du mot X (autrement dit, probabilité d'occurrence du mot X).

1. http://www.lexically.net/wordsmith
2. http://www.antlab.sci.waseda.ac.jp/software.html
3. http://cwb.sourceforge.net

On se trouve ici dans un modèle dit *de vraisemblance maximale* (on évalue à quel point l'hypothèse de corrélation est vraisemblable) dans lequel la probabilité d'un événement est estimée par sa fréquence. C'est-à-dire, la probabilité obtenue de co-occurrence de deux mots X et Y, $P(X,Y)$, est postulée égale à la fréquence relative de co-occurrence de ces deux mots dans un corpus donné $Freq(X\&Y)$. La probabilité $P(X)$ du mot X est égale à sa fréquence relative dans le même corpus $Freq(X)$. La probabilité espérée est calculée ici à partir des probabilités des mots X et Y. En termes de fréquences des mots dans le corpus, nous avons alors :

$$IM(X,Y) = \log_2 \frac{Freq(X\&Y)}{Freq(X)Freq(Y)} \tag{3.3}$$

Plus la valeur IM obtenue est élevée, plus la corrélation entre les mots X et Y est forte.

La mesure *log-likelihood* est calculée ainsi (Rapp, 1999) :

$$-2\log\lambda = \sum_{i,j\in\{1,2\}} k_{i,j}\log\frac{k_{ij}N}{C_iR_j} = k_{11}\log\frac{k_{11}N}{C_1R_1} + k_{12}\log\frac{k_{12}N}{C_1R_2} + k_{21}\log\frac{k_{21}N}{C_2R_1} + k_{22}\log\frac{k_{22}N}{C_2R_2} \tag{3.4}$$

où

- $C_1 = k_{11} + k_{12}$
- $C_2 = k_{21} + k_{22}$
- $R_1 = k_{11} + k_{21}$
- $R_2 = k_{12} + k_{22}$
- $N = k_{11} + k_{12} + k_{21} + k_{22}$

et

- k_{11} renvoie à la fréquence de co-occurrence des mots X et Y (égale à $Freq(X\&Y)$)
- k_{12} renvoie à la fréquence du mot X lorsqu'il apparaît seul (pas de co-occurrence avec Y)
- k_{21} renvoie à la fréquence du mot Y lorsqu'il apparaît seul (pas de co-occurrence avec X)
- k_{22} renvoie à la taille du corpus en enlevant les fréquences des mots X et Y.

Contrairement à *l'information mutuelle*, cette mesure prend en compte non seulement la probabilité d'apparition d'un mot dans le corpus, mais la probabilité de son apparition en dehors des co-occurrences avec le deuxième mot étudié. Plus la valeur obtenue est élevée, plus la corrélation entre les deux mots est forte.

En linguistique de corpus, ces mesures sont exploitées pour identifier des collocations, c'est-à-dire des groupes de mots « qui ont tendance à apparaître à proximité l'un de l'autre significativement plus souvent qu'on pourrait le prédire en se basant sur la fréquence d'occurrence de chaque mot pris individuellement »[4]. (Kilgarriff, 1992).

Plusieurs implémentations de ces mesures ou d'autres mesures d'association dans les applications du TAL permettent de révéler dans un corpus les syntagmes avec un degré élevé d'unité, parmi tous les segments appelés *clusters* ou *n-grammes*.

Néanmoins, en examinant la sortie d'un extracteur, AntConc[5], (cf. figure 3.1), on s'aperçoit de nombreux problèmes :

1. Le nombre de mots que doivent comporter les syntagmes analysés n'est pas fixe (cf. *machine synchrone* vs. *gaz à effet de serre*). Dans les concordanciers, ce nombre - la taille de la fenêtre - est d'habitude passé en paramètre ;
2. Le seuil à partir duquel nous pouvons constater un haut degré d'unité varie beaucoup en fonction de la taille du corpus (L'Homme, 2004) et ne peut pas être défini a priori ;

4. « A collocation is a group of two or more words which are to be found in proximity to each other significantly more often than one would predict, given the frequency of occurrence of each word taken individually. » (Kilgarriff, 1992, p. 29).

5. AntConc (Version 3.2.2). Tokyo, Japan : Waseda University, http://www.antlab.sci.waseda.ac.jp

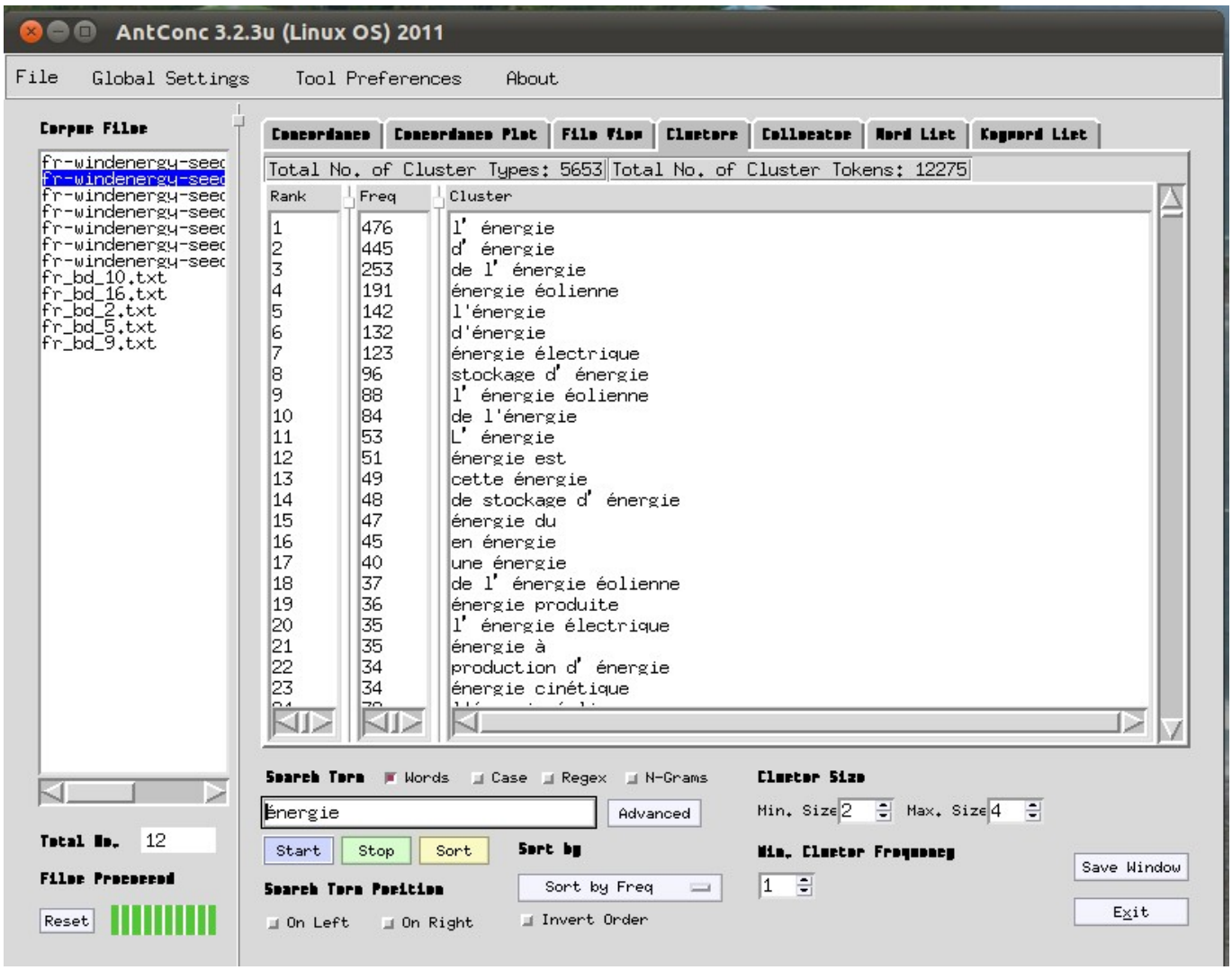

FIGURE 3.1 – Extrait de la sortie du concordancier AntConc : les clusters du mot *énergie* dans le corpus d'énergie éolienne

3. Les sorties d'outils sont « bruitées » par des mots usuels (*un appel*, *téléphoner à*, etc.) ;
4. Les composants d'un terme complexe sont parfois distants (EN *wind and solar energy* 'énergie éolienne et solaire' - variante de *wind energy* 'énergie éolienne' ;
5. Parmi les syntagmes identifiés, il faut distinguer les termes polylexicaux (*énergie cinétique*) des collocations des termes simples qui ne sont pas des termes polylexicaux (*énergie produite*).

On peut échapper à certaines de ces difficultés en exploitant des techniques linguistiques, notamment celles des patrons morphosyntaxiques et celles des frontières lexicales.

Les patrons morphosyntaxiques sont des séquences typiques de catégories grammaticales. Pour l'extraction terminologique, on examine en premier lieu les syntagmes nominaux. Les chercheurs tiennent pour acquis que la plupart des termes complexes dans les textes spécialisés sont des groupes nominaux (L'Homme, 2004). Ainsi, pour extraire les candidats termes, il faut extraire les patrons typiques. Les patrons varient selon la langue : les patrons principaux en français sont [N N], [N PREP N] et [N ADJ], tandis qu'en japonais ce sont [N N], [N SUF ADJ], [ADJ N] et [PREF N] (Morin et al., 2007). La séquence [N ADJ] en français correspond à la séquence [ADJ N] en russe et souvent à la séquence [N N] en anglais : FR *énergie*$_N$ *éolienne*$_{ADJ}$ - RU ветровая$_{ADJ}$ энергия$_N$ - EN *wind*$_N$ *energy*$_N$. Les patrons peuvent aider à distinguer les termes polylexicaux des collocations d'un terme qui ne sont pas des termes, par exemple (en français) l'absence d'article est caractéristique des termes polylexicaux (*moyeu de roue*), vs. collocation qui

n'a pas de statut terminologique : *hauteur du moyeu*. Afin d'identifier les patrons susceptibles de former des termes dans une langue, deux méthodes sont possibles : (1) l'expertise manuelle par les spécialistes de la langue (Weller et al., 2011) ou (2) l'acquisition automatique des patrons à partir d'un corpus d'apprentissage annoté manuellement (Guégan et De Loupy, 2011).

L'autre technique abordée ici, « les frontières de termes », a été proposée par Bourigault (1994) et implémentée dans l'outil LEXTER. Elle consiste à définir les frontières éventuelles entre les unités terminologiques. Les éléments qui forment les frontières sont :

— les signes de ponctuation ;
— les verbes (cela découle du principe que la large majorité des termes complexes sont des groupes nominaux) ;
— les pronoms ;
— les déterminants précédés d'un verbe ou d'un signe de ponctuation ;
— etc.

Voici un exemple du découpage du texte par LEXTER tiré de (Bourigault et al., 1996) :

> (L')opérateur [passe en] recirculation directe sur puisard [grâce à (la)] vanne manuelle d'isolement d'enceinte [qui est équipée (du)] clapet de sécurité rapide.

Les mots entre crochets forment des frontières entre les candidats termes complexes.

Par rapport aux techniques statistiques, les techniques linguistiques permettent d'effectuer une analyse plus fine et d'éviter plusieurs cas de mauvais découpage du terme (e.g. *téléphoner à*). Par contre, elles exigent des corpus étiquetés (avec des parties du discours) et des règles différentes en fonction de la langue (même si ces règles sont généralement assez simples). Il est possible de combiner des techniques statistiques et linguistiques. Des exemples d'outils qui exploitent à la fois le calcul de fréquence et l'extraction de patrons morphosyntaxiques sont Acabit et TermSuite.

3.1.3 Extraction des termes en utilisant des ressources sémantiques

D'autres méthodes d'extraction des termes apparaissent avec la disponibilité de ressources encyclopédiques en ligne. La source largement utilisée est Wikipédia [6], la plus grande encyclopédie sur le web. Nous allons citer ici quelques exemples d'utilisation de Wikipédia pour l'extraction des termes monolexicaux ainsi que polylexicaux.

L'approche « *top down* » ('de haut en bas') (Vivaldi et Rodríguez, 2010a) consiste à choisir manuellement une catégorie dans l'encyclopédie correspondant à un domaine (par exemple, *médecine*), et ensuite à extraire automatiquement tous les termes attachés à cette catégorie. Les termes, dans ce cas, sont les titres de pages et les catégories liées dans l'encyclopédie à la catégorie choisie, et filtrées par le système.

L'approche « *bottom up* » ('de bas en haut') (Vivaldi et Rodríguez, 2010b) exige aussi une catégorie prédéfinie. Toutefois, l'analyse commence non par cette catégorie, mais par une liste des candidats termes issus d'un texte spécialisé (donc « du bas »). Pour chaque candidat, on cherche la page de Wikipédia correspondante et on mesure son coefficient d'appartenance au domaine (« *Domain Coefficient* »), i.e. à la catégorie choisie, en fonction de la quantité et de la longueur des chemins remontant de cette page vers la catégorie choisie.

Vivaldi et Rodríguez (2012) combinent les deux approches. Pour choisir la (les) catégorie(s) de Wikipédia correspondant au domaine d'intérêt, ils utilisent d'autres ressources sémantiques, telles que Magnini's Domain Codes (Magnini et Cavaglià, 2000) et WordNet [7]. En utilisant des classes de ces taxonomies, ils établissent une liste préalable des catégories de Wikipédia, ainsi qu'une liste préalable des pages de Wikipédia, relatives au domaine choisi. Ensuite, ils appliquent l'approche « *bottom up* » à partir de cette liste

6. http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal
7. http://wordnet.princeton.edu/

de pages pour filtrer la liste des catégories, et enfin l'approche « *top down* » pour étendre cette dernière liste. Dans l'hypothèse où les termes corrects du domaine sont ceux qui apparaissent dans ces deux listes simultanément, ils continuent à générer des listes modifiées de catégories et de pages jusqu'à ce que ces deux listes convergent.

Les candidats termes trouvés avec de telles méthodes sont des expressions bien formées : les problèmes de découpage de terme et de degré d'unité ne se posent pas comme avec l'utilisation de corpus de textes. Toutefois, le problème du degré de spécialisation demeure : les ressources de type Wikipédia sont des ressources libres de nature encyclopédique et non spécialisée dans un domaine donné ou pour une application précise. Le statut terminologique de certains candidats termes extraits à partir de telles sources prête à controverse.

3.2 Extraction terminologique multilingue

L'extraction terminologique bilingue (ou multilingue) comporte deux aspects : (1) l'extraction des termes et (2) l'alignement des termes, c'est-à-dire l'établissement des équivalences entre les termes de deux langues abordées. Le deuxième aspect rejoint dans une certaine mesure la traduction automatique. L'extraction multilingue peut être accomplie à partir de corpus parallèles ou, étant donné la rareté de ces ressources, à partir de corpus comparables.

3.2.1 Extraction terminologique à partir de corpus parallèles

Pour l'alignement des termes, et plus largement pour la traduction automatique, on a souvent besoin que les corpus soient alignés, c'est-à-dire que chaque unité du texte source (souvent une phrase ou un paragraphe) soit mise en correspondance avec une unité du texte cible.

Les corpus alignés existants possèdent généralement déjà un alignement réalisé au niveau des phrases. Pour l'extraction bilingue comprenant l'alignement lexical, deux schémas majeurs sont possibles : (1) de l'extraction monolingue vers l'alignement des termes : les termes candidats sont d'abord extraits indépendamment pour chaque langue (ou pour une des deux langues), et ensuite des équivalences entre les termes sont établies en s'appuyant sur l'ordre des mots, sur le formatage (italique, gras, etc.) ou sur l'alignement mot à mot fait préalablement au sein des phrases alignées ; (2) de l'alignement des syntagmes vers l'extraction terminologique : les phrases sont segmentées en syntagmes non récursifs (« *chunks* »), et les syntagmes correspondants sont alignés entre les deux langues. Enfin, seuls les syntagmes dont le statut terminologique est attesté (en fonction de leur fréquence dans le corpus) sont retenus.

La première approche est illustrée par les travaux de Gaussier (1998) et de Vintar (2010) qui nécessitent un alignement préalable mot à mot. Pour l'effectuer, Gaussier (1998) introduit une méthode de graphes construits en se basant sur des mesures de probabilité d'association de deux mots. Ensuite, il extrait des termes du corpus source et recherche leurs équivalents dans le corpus cible grâce aux graphes construits qui lient les mots d'une langue à l'autre. Un travail plus récent de Vintar (2010) porte sur le couple de langues anglais-slovène. L'auteur extrait des termes simples et complexes de chaque sous-corpus (anglais et slovène), indépendamment. Ensuite les termes extraits des deux langues sont alignés en utilisant le lexique bilingue construit préalablement à partir du corpus parallèle, à l'aide de l'outil d'alignement Twente [8].

Un exemple de cette deuxième approche est le travail de Lefever et al. (2009). Au cours de la première étape, les auteurs font l'alignement des corpus parallèles au niveau de segments inférieurs à la phrase (il peut s'agir des signes de ponctuation, des mots, des syntagmes). Ils repèrent d'abord les segments courts (mots graphiques, signes de ponctuation) de la langue source qui sont en correspondance « sûre » avec

8. http://wwwhome.cs.utwente.nl/~irgroup/align/download.html

ceux de la langue cible (« *anchor chunks* ») à l'aide du dictionnaire. Ensuite, ils alignent des segments plus larges (potentiellement des termes complexes) sur la base de certaines règles (ordre des mots, etc.). Dans un deuxième temps, ils filtrent les candidats termes simples extraits par degré de spécificité (en utilisant la mesure de *log-likelihood*) et les candidats termes complexes par degré d'unité (en utilisant la mesure d'*espérance mutuelle*) pour retenir les vrais termes. Ils ont mené des expérimentations sur trois couples de langues (français-anglais, français-italien et français-hollandais) sur le corpus automobile.

3.2.2 Extraction terminologique à partir de corpus comparables

Les corpus parallèles de taille importante sont des ressources rares (d'autant plus pour les corpus parallèles spécialisés) car leur construction est un processus coûteux en termes d'efforts humains. C'est pourquoi certains chercheurs recourent à l'exploitation des corpus comparables pour l'extraction terminologique bilingue. Nous examinons ici deux approches principales pour l'extraction et l'alignement terminologique, à savoir l'approche distributionnelle et l'approche compositionnelle.

Approche distributionnelle

Dans les corpus parallèles, la co-occurrence des mots X et Y de la langue source est aussi fréquente que la co-occurrence de leurs traductions X' et Y' dans la langue cible. Rapp (1995) a démontré que cette affirmation est valable également pour les corpus comparables. Par exemple, si dans un corpus anglais on trouve une co-occurrence régulière des mots *blue* ('bleu') et *sky* ('ciel'), alors dans un corpus comparable allemand on trouvera une co-occurrence régulière de leurs analogues *blau* et *Himmel*. Pour vérifier cette hypothèse, il a comparé les matrices de co-occurrence de six mots (*bleu, vert, plante, école, ciel, professeur*) tirées de corpus comparables anglais et allemand. Il a constaté une forte corrélation entre les co-occurrences régulières dans ces matrices.

Dans un autre travail, Rapp (1999) développe ses réflexions sur le sujet en proposant une méthode de vecteur de contexte. Il utilise deux corpus non parallèles de langues anglaise et allemande, et un petit dictionnaire bilingue (environ 16000 mots) afin de déterminer les traductions des mots qui ne font pas partie de ce dictionnaire. Pour atteindre ce but, il construit d'abord une matrice de co-occurrences pour la langue cible, dans laquelle les lignes correspondent à tous les mots différents du corpus lemmatisé, et les colonnes à tous les mots de cette langue présents dans le dictionnaire du départ. Les co-occurrences dans la matrice sont calculées en utilisant la mesure *log-likelihood*.

Ensuite, pour traduire un mot-pivot du corpus de la langue source (n'appartenant pas au dictionnaire), l'auteur construit le vecteur de contextes de ce mot dans le corpus source (incluant tous les mots qui co-occurrent avec le mot-pivot dans une fenêtre de taille 3). Il cherche les traductions des mots de ce vecteur dans le petit dictionnaire de départ, et les traductions trouvées forment un nouveau vecteur (les mots dont les traductions ne sont pas trouvées sont écartés du vecteur). Il change l'ordre des mots dans le nouveau vecteur pour l'ajuster à celui de la matrice initiale. Enfin, il compare ce vecteur (constitué maintenant de mots en langue cible) avec tous les vecteurs dans la matrice afin de choisir le vecteur le plus proche. Ce mot du corpus cible dont le vecteur est choisi est admis comme une traduction du mot-pivot du corpus source. Cette méthode a été utilisée comme base pour plusieurs recherches en extraction terminologique bilingue.

Approche compositionnelle

La méthode distributionnelle est généralement utilisée pour l'alignement des termes simples. Pour les termes complexes (monolexicaux et polylexicaux), l'approche compositionnelle est exploitée.

Cette méthode a été introduite par Grefenstette (1999) pour la traduction automatique à partir de grands volumes de données langagières (l'auteur utilise pour ces expérimentations le web en tant que corpus),

initialement pour des expressions de la langue générale et non des domaines de spécialité. L'auteur cherche à traduire en anglais les composés monolexicaux allemands et les expressions polylexicales espagnoles. Il utilise un dictionnaire bilingue pour sélectionner les expressions complexes à traduire. Les critères sont : (1) la traduction d'un ensemble peut être trouvée en combinant les traductions des parties et (2) ce processus doit aboutir à plusieurs candidats pour la traduction en anglais. Ensuite il génère tous les candidats de traductions possibles en traduisant chaque composant séparément (avec l'aide du même dictionnaire). Enfin, il lance des requêtes sur ces candidats dans un moteur de recherche (AltaVista), et il calcule la fréquence de chaque candidat. Le candidat avec le meilleur score est considéré comme la bonne traduction.

Cette méthode utilisant le web pour valider les candidats de traductions apporte les traductions les plus communes, ce qui n'est pas un problème pour la traduction des textes de la langue générale, contrairement aux langues spécialisées, pour lesquelles une traduction plus rare peut s'avérer meilleure. Cette méthode a été adaptée à l'alignement des termes, en exploitant des corpus spécialisés au lieu du web, à savoir des corpus comparables (Morin et al., 2007) ainsi que des corpus parallèles (Vintar, 2010).

Il est possible de combiner deux approches en utilisant la méthode compositionnelle pour la traduction des termes complexes, et la méthode distributionnelle pour la traduction des termes simples (Morin et al., 2007) et éventuellement pour la traduction d'un des composants du terme complexe pour lequel la traduction n'est pas présente dans le lexique bilingue utilisé (Morin et Daille, 2012). Une telle approche « mixte » est implémentée dans le logiciel TermSuite (version 1.4).

3.3 Bilan

Dans ce chapitre, nous avons décrit les méthodes principales d'extractions automatiques monolingue et bilingue. Nous avons observé que souvent des techniques différentes sont appliquées pour extraire les termes simples ou les termes complexes. Nous avons abordé le fonctionnement, les possibilités et les limites de ces méthodes, mais nous n'avons pas comparé leur efficacité. L'évaluation des résultats de l'extraction des termes sera traitée dans le chapitre suivant.

4

Évaluation des résultats des extracteurs terminologiques

Ces dernières années, des avancées considérables dans le domaine de l'acquisition automatique de terminologie ont été réalisées (Mustafa El Hadi et al., 2006). Malgré cela, les extracteurs existants sont encore loin d'être parfaits. Tous les systèmes d'extraction terminologique présentent des avantages mais aussi des inconvénients. Certains problèmes relèvent de la particularité de cette tâche et sont incontournables (en tout cas, jusqu'à présent), d'autres sont propres à la méthode utilisée. Pour évaluer les systèmes et les comparer entre eux, des procédures rigoureuses et applicables à des outils différents doivent être élaborées.

Dans ce chapitre, nous résumons les stratégies d'évaluation couramment utilisées pour l'extraction terminologique monolingue et bilingue (section 4.1). Nous nous concentrons sur la stratégie qui recourt à une liste de termes de référence. Nous rapportons notre expérience de la construction de telles listes et les difficultés rencontrées (section 4.2). Nous terminons en analysant les résultats de l'évaluation d'un extracteur (TermSuite) mise en place avec l'utilisation des listes construites (section 4.3).

4.1 Stratégies d'évaluation des extracteurs terminologiques

Les deux stratégies les plus courantes d'évaluation des résultats fournis par un extracteur sont l'expertise humaine et la comparaison avec un référentiel (liste de référence).

Évaluation de l'extraction terminologique monolingue. Dans un cadre monolingue, la première stratégie d'évaluation consiste à faire analyser la sortie d'un extracteur (généralement la sortie est une liste de termes candidats) par un expert du domaine afin de marquer pour chaque candidat s'il s'agit d'un terme du domaine ou non. La deuxième stratégie consiste à comparer (de manière automatique) la sortie d'un outil avec une liste de référence préétablie. Par liste de référence (« *reference term list* », par la suite **RTL**), nous entendons une liste des termes les plus importants et caractéristiques du domaine donné. Les deux stratégies ont leurs points forts et points faibles. Nous résumerons les points clés mentionnés à ce sujet dans les travaux de Mustafa El Hadi et al. (2004, 2006) dans le tableau 4.1.

TABLE 4.1 – Comparaison des deux stratégies d'évaluation des extracteurs terminologiques.

Expertise humaine		Liste de référence	
+	applicable aux logiciels de tous types et pour tous domaines	+	peut être réutilisée dans plusieurs évaluations
-	coûteuse en termes d'efforts et de temps (pour chaque système ou en cas de nouvelle configuration, l'évaluation est à refaire)	+	permet d'évaluer le silence (les termes importants qui ne sont pas trouvés par le système évalué)
-	le degré d'accord entre experts sur le statut terminologique est assez bas	-	les termes importants qui ne sont pas dans la RTL ne seront pas comptés comme corrects

Évaluation de l'extraction terminologique bilingue. La sortie d'un extracteur des termes bilingues est une liste d'alignements entre les termes de la langue source (TS) et les termes de la langue cible (TC). Les alignements peuvent être univoques (1 TS correspond à 1 TC) ou multiples (1 TS correspond à n TC). Dans l'évaluation du premier type (expertise humaine), un expert analyse la qualité des alignements proposés par le système. Dans l'évaluation du deuxième type, les alignements proposés sont comparés avec une liste de référence. La **RTL bilingue** contient des termes caractéristiques du domaine de la langue source mis en correspondance avec leurs équivalents de la langue cible.

Si l'évaluation porte sur toute la chaîne de l'extraction (monolingue et ensuite bilingue), la qualité des alignements doit être analysée en tenant compte de trois aspects : le statut terminologique du TS, le statut terminologique du TC et l'équivalence entre les deux (Vintar, 2010). Sinon, l'extraction bilingue peut être évaluée indépendamment de l'extraction monolingue. Dans ce cas, l'ensemble des TS et TC extraits est filtré manuellement de manière à n'y conserver que des termes spécifiques pour le domaine donné, et l'évaluation de l'extraction bilingue est équivalente à l'évaluation uniquement de la qualité des alignements établis. Ce scénario est réalisé dans une campagne d'évaluation de l'extraction français-japonais à partir des corpus comparables (Morin et al., 2007). Dans cette campagne, deux RTLs ont été utilisées : RTL1 comprenant uniquement des termes simples FR dont les traductions sont également des termes simples JP, et RTL2 comprenant des termes simples et complexes FR dont les traductions sont des termes simples ou complexes JP. Les auteurs ont pointé la différence entre les structures linguistiques dans les deux langues : la plupart des termes français sont polylexicaux, tandis que les termes japonais tendent vers la composition. Cela explique l'écart considérable entre les résultats obtenus : le taux de termes FR ayant un alignement correct JP parmi les Top N meilleurs candidats dans RTL1 s'est avéré de 49 % pour le Top 10 et de 52 % pour le Top 20 (sur les corpus contenant uniquement des textes scientifiques), tandis que dans RTL2 ce taux a été nettement moins élevé, 18 % pour le Top 10 et 25 % pour le Top 20. Nous pouvons donc rajouter encore un point faible de l'évaluation utilisant la liste de référence : le choix des termes dans la liste influence les résultats. Malgré cela, le principal avantage de cette technique reste valable dans le contexte de l'extraction bilingue, à savoir la réutilisation potentielle d'une RTL pour évaluer plusieurs systèmes ou configurations de paramètres, tandis que l'expertise humaine doit être refaite à chaque fois.

Chaudiron (2001) propose une autre technique d'évaluation, l'évaluation par les utilisateurs finaux du système du point de vue de leur satisfaction et de l'adaptabilité du système aux tâches pour lesquelles il n'était pas conçu. Cette technique se rapproche certainement le plus de l'évaluation d'un système dans les conditions réelles d'utilisation, et elle peut être appliquée à l'extraction monolingue ainsi que bilingue. Toutefois elle partage un inconvénient important avec l'expertise humaine : c'est une procédure coûteuse en termes d'efforts et de temps, et elle est à refaire pour la sortie de chaque système ou chaque configuration d'un système.

Dans le cadre du projet TTC [1], nous avons eu l'occasion de participer à l'évaluation des outils d'extraction terminologique élaborés au cours du projet. L'évaluation a été effectuée en utilisant des listes terminologiques de référence car l'objectif était d'évaluer et de comparer plusieurs systèmes d'extraction sur sept langues, et l'expertise humaine aurait été difficile à mettre en place. Nous discutons par la suite de notre expérience de construction des listes de référence pour cette évaluation.

4.2 Construction des listes de référence

Pour pouvoir évaluer l'extraction terminologique monolingue et bilingue, nous avons constitué manuellement les RTLs monolingues pour l'anglais et le français, ainsi que les RTLs bilingues anglais-français et français-russe pour les deux domaines : l'énergie éolienne et les technologies mobiles. Nos partenaires TTC ont effectué cette tâche pour les autres langues du projet, c'est-à-dire l'allemand, l'espagnol, le letton, le russe et le chinois. Les RTLs ont été construites à partir de corpus comparables spécialisés. Les consignes pour la construction des RTLs ont été définies au préalable par les membres du projet, dans un objectif d'homogénéisation des RTLs, et elles ont été partiellement revues en cours de réalisation.

4.2.1 Consignes pour la construction des RTLs

La taille des RTLs était définie par avance : 130 termes par langue et par domaine pour les RTLs monolingues, et 100 paires de termes pour les RTLs bilingues. Les outils d'extraction terminologique que le projet visait à évaluer seraient testés sur des corpus de textes. Par conséquent, pour assurer la possibilité de trouver tous les termes d'une RTL pendant l'évaluation, les RTLs ne devaient inclure que des termes appartenant à ces corpus. Nous avons utilisé des corpus spécialisés comparables (pour les statistiques se reporter au tableau B.2).

Les RTLs devaient inclure des termes monolexicaux (SWT) ainsi que des termes polylexicaux (MWT), des termes de base ainsi que leurs variantes. Nous avons considéré comme variantes toutes les unités lexicales synonymiques ou liées sémantiquement avec le terme de base (Daille, 2005), à savoir des variantes graphiques (EN *offshore - off-shore*), morphologiques (*synchrone - asynchrone*), syntaxiques (*énergie éolienne - énergie électrique éolienne*) et des variantes paradigmatiques formées par substitution d'un des éléments du terme complexe par un synonyme (*ferme éolienne - parc éolien*). D'autres caractéristiques utiles pour l'évaluation ont été ajoutées pour chaque terme, notamment : la catégorie grammaticale, l'annotation morphologique (avec le jeu d'étiquettes MULTEXT [2]), toutes les formes fléchies rencontrées dans le corpus, les fréquences du terme et de ses variantes, ses collocations fréquentes, et l'origine pour les termes dits néoclassiques (i.e. d'origine grec ou latine) ou les emprunts (cf. exemple dans le tableau 4.2).

La répartition des termes de types différents était fixée a priori pour homogénéiser les RTLs : 20 % de termes simples (monolexicaux), 20 % de termes complexes monolexicaux et 60 % de termes complexes polylexicaux. Cependant, au cours de leur réalisation, cette condition s'est avérée difficile à respecter en raison de la différence typologique selon les langues.

Puisque la fréquence du candidat est un facteur important que les systèmes d'extraction prennent en compte, une fréquence minimale d'occurrences d'un terme de RTL dans le corpus a été définie (arbitrairement) : de 10 pour les SWT et de 5 pour les MWT.

1. Terminology Extraction, Translation Tools and Comparable Corpora http://www.ttc-project.eu
2. http://aune.lpl.univ-aix.fr/projects/multext, date de consultation 24.05.12

TABLE 4.2 – Extrait de la liste de référence EN-FR, pour le domaine de l'énergie éolienne.

	Source language	Target language
term lemma	sustainable energy	énergie durable
SWT/MWT	MWT	MWT
pattern	A N	N A
morph. tag	A—s- Nc-s-	Ncfs- A—ms-
origin		
inflected forms (IF)		énergies durables
frequency	39	5
most frequent IF (mfIF)	sustainable energy	énergie durable
frequency of mfIF	39	4
variant	green energy	
frequency	16	
variant type	paradigmatic	
synonym	yes	
collocations		

4.2.2 Méthode de construction des RTLs

Une liste de référence peut être construite à partir d'une ressource terminologique préexistante (un glossaire du domaine, un thésaurus) ou directement à partir d'un corpus spécialisé sur lequel l'outil sera testé par un expert du domaine (Mustafa El Hadi et al., 2004). Dans le cadre du projet, la deuxième option a été choisie, à ceci près que ce n'étaient pas des experts du domaine qui construisaient les RTLs. C'est pourquoi nous avons eu recours à la méthode semi-automatique pour la construction des RTLs, i.e. avec l'utilisation de concordanciers ou d'outils terminologiques disponibles (AntConc, Acabit).

RTLs monolingues. Pour retenir des termes monolexicaux pour une RTL, nous avons appliqué le concordancier AntConc au corpus spécialisé afin de classer les mots du corpus par spécificité (cf. formule 6.6). Nous avons examiné les premiers résultats de cette liste (jusqu'à 2000-3000 candidats) pour sélectionner manuellement des termes du domaine. Parfois pour prendre la décision finale de conserver ou non un terme, nous avons analysé les contextes d'apparition du terme-candidat dans les corpus ou sur le web.

Pour calculer la spécificité, nous avons besoin de corpus de référence. Dans un but d'homogénéisation, des corpus journalistiques ont été choisis comme corpus de référence, parce qu'ils sont disponibles pour toutes les langues TTC. Même s'il s'agit de corpus d'un genre particulier, ils sont suffisamment généraux du point de vue du sujet par rapport à nos domaines étudiés. Les statistiques des corpus sont présentées dans le tableau B.2.

Pour sélectionner les termes polylexicaux, nous avons exploité la sortie de l'extracteur de termes Acabit (extracteur de termes polylexicaux), et nous l'avons complété en utilisant des co-occurrences fréquentes extraites avec AntConc. Nous avons examiné les co-occurrences des termes monolexicaux retenus s'avérant beaucoup plus fréquentes dans le corpus spécialisé que dans le corpus général, mais pas suffisamment spécifiques pour être rajoutées individuellement dans la RTL comme termes simples. Par exemple, les adjectifs français *vertical* et *horizontal* sont très fréquents dans le corpus de l'énergie éolienne, mais on ne peut pas les considérer comme des termes caractéristiques du domaine. Cependant, leurs co-occurrences - *axe vertical* et *axe horizontal* - sont des termes polylexicaux fiables caractéristiques d'une turbine éolienne.

TABLE 4.3 – Tailles des RTLs monolingues construites.

	Technologies mobiles		Énergie éolienne	
	termes	variantes	termes	variantes
EN	140	17	130	40
FR	131	19	126	80

Nous avons également appliqué des critères linguistiques de terme (L'Homme, 2004, p. 65-66) :

— La parenté morphologique : les dérivés d'un terme sont des termes : *rotor – rotorique*, *pompe – pompage* ;

— Les relations paradigmatiques : les candidats liés à un terme par des relations paradigmatiques diverses (synonymie, antonymie, hyponymie, hyperonymie, etc.) sont très probablement des termes. Des indices de telles relations peuvent être trouvés dans les définitions terminologiques. Par exemple, la définition du terme *hélice* dans le Grand Dictionnaire Terminologique - *« partie du rotor de l'éolienne constituée de l'ensemble des pales et du moyeu »* - se réfère aux termes *rotor*, *éolienne*, *pale*, *moyeu*, alors *hélice* peut aussi être admis comme un terme.

— Les contextes particuliers : l'emploi terminologique d'une unité lexicale conduit souvent à l'apparition de collocations spécialisées différentes des collocations typiques de cette unité dans la langue générale : *arbre* dans sa signification *« tige qui tourne au centre d'une génératrice ou d'un rotor d'éolienne et qui achemine l'énergie »* [3] dans les textes consacrés à l'énergie éolienne apparaît dans une collocation *arbre lent* n'existant pas dans la langue générale.

De plus, nous avons remarqué que les mots d'origine étrangère (surtout les mots anglais) sont très fréquents dans les textes spécialisés. Ainsi, les mots anglais *wind* 'vent', *energy* 'énergie', *power* 'puissance', *speed* 'vélocité' apparaissent en tête de la liste de fréquence française issue de notre corpus. Les raisons sont les suivantes : (1) certains termes récents n'ont pas encore de traduction conventionnelle ; (2) les équivalents anglais des termes importants sont souvent donnés dans les textes spécialisés ; (3) les textes incluent parfois une partie en langue étrangère (le résumé d'articles scientifiques et de thèses, etc.) Les traductions de ces mots étrangers sont potentiellement de bons termes candidats.

A la fin, nous avons vérifié si les termes candidats retenus étaient recensés dans une des grandes banques terminologiques telles que TERMIUM Plus [4], Grand Dictionnaire Terminologique [5], IATE [6], EuroTermBank [7] etc. Pour identifier les variantes des termes, nous avons exploité la sortie d'Acabit.

Les tailles exactes des RTLs monolingues sont données dans le tableau 4.3.

RTLs bilingues. Les RTLs bilingues ont été produites sur la base des RTLs monolingues. La taille d'une RTL monolingue avait été fixée à 130 afin d'assurer la présence de 100 termes en commun entre deux listes monolingues afin de constituer la liste bilingue. Cependant, cette marge de 30 % s'est révélée insuffisante pour obtenir 100 couples de termes. Les raisons sont multiples :

— Certains néonymes anglais entrent en usage dans d'autres langues, les équivalents natifs n'étant pas encore introduits ou normalisés (ex. du domaine des communications mobiles : *bluetooth, smartphone, roaming, streaming*, etc.).

— La fréquence varie beaucoup en fonction de la langue : l'équivalent d'un terme fréquent dans une langue peut n'apparaître qu'une seule fois dans le corpus d'une autre langue, et par conséquent, ne répondra pas à nos critères de choix de termes (plus précisément à celui de fréquence minimale) ;

3. TERMIUM Plus
4. http://www.btb.termiumplus.gc.ca/
5. http://www.gdt.oqlf.gouv.qc.ca
6. iate.europa.eu/
7. http://www.eurotermbank.com/

TABLE 4.4 – Tailles des RTLs bilingues construites (nombre de termes).

	Technologies mobiles	Énergie éolienne
EN-FR	100	103
FR-RU	100	100

— Les RTLs monolingues sont limitées en taille et ne sont pas exhaustives, certains termes peuvent donc apparaître dans la RTL d'une langue source, mais pas dans celle d'une langue cible, même s'ils sont présents et suffisamment fréquents dans le corpus de la langue cible.
— La traduction d'un terme spécialisé de la langue source peut être polysémique en langue cible et largement employée dans la langue générale, de ce fait elle n'a pas été extraite comme terme en langue cible : le terme FR *portance* se traduit en EN comme *lift* ;

Pour résoudre ce problème, nous avons été obligée de compléter les RTLs bilingues par des termes présents dans les deux corpus, mais pas dans les deux RTLs monolingues. Nous avons donc cherché les traductions de certains termes de la RTL d'une langue source dans le corpus de la langue cible. Les tailles des RTLs bilingues EN-FR sont données dans le tableau 4.4.

4.2.3 Problèmes rencontrés au cours de la construction des RTLs

Pendant la construction des listes de référence, nous avons rencontré un certain nombre de difficultés récurrentes dans les pratiques terminologiques, par conséquent, avec d'autres membres du projet, nous avons été obligés de retenir des conventions opérationnelles pour les contourner ou les minimiser.

RTLs monolingues.

1. **Interférence des domaines.** Une difficulté majeure est de définir dans quels cas on assimile une unité lexicale à un terme. La spécialité par rapport à un domaine, le critère principal du statut terminologique, est souvent mesurée en utilisant la fréquence du candidat dans un corpus du domaine. Toutefois, dans tous les corpus exploités, nous avons trouvé des termes très fréquents issus de domaines adjacents comme *énergie solaire* ou *énergie nucléaire* dans le corpus de l'énergie éolienne, ou bien les termes du domaine des réseaux dans les corpus des technologies mobiles. Pour les termes les plus importants provenant des domaines adjacents, nous avons opté pour leur inclusion dans les RTLs.
2. **Contrainte de fréquence minimale.** Plus la taille du corpus spécialisé est grande, plus le lien entre la fréquence et la spécificité d'une unité est fiable. Les corpus que nous avons utilisés ne sont pas très étendus, et parfois les termes pertinents du domaine ont une fréquence basse dans un des corpus, inférieure à la fréquence minimale convenue pour les termes dans les RTLs. Nous avons gardé la contrainte initiale d'une fréquence minimale afin d'assurer un nombre suffisant de contextes pour l'extraction automatique.
3. **Répartition des types de termes définie a priori.** Une autre contrainte était les proportions de termes mono- et polylexicaux définies pour toutes les langues. En réalité, ces proportions dépendent de la langue. Par exemple, en allemand le taux de termes monolexicaux est beaucoup plus élevé qu'en EN ou FR à cause des composés monolexicaux. La formation de nouveaux mots par composition est très productive en DE, et beaucoup de termes polylexicaux anglais et français doivent être traduits correctement en allemand avec des termes monolexicaux : FR *pale du rotor* - DE *Rotorblatt*, FR *surface du rotor* - DE *Rotorfläche*, etc. Il a donc été décidé de renoncer à cette distribution initiale pour privilégier la réalité de chaque langue dans les RTLs monolingues. De même, les proportions des termes néoclassiques, les patrons syntaxiques typiques formant les termes et leur productivité,

ainsi que les types de variantes terminologiques, varient d'une langue à une autre, aucune répartition universelle dans les RTLs ne peut être envisagée.

4. **Abréviations et sigles.** D'après les consignes initiales, les abréviations et les sigles n'étaient pas considérés comme des entrées terminologiques indépendantes, mais comme des variantes de leurs formes développées. Par conséquent, nous ne devions pas les inclure dans les RTLs. En réalité, les abréviations et les sigles sont très fréquentes dans les langues de spécialité, plus que leurs formes développées. Nous l'avons constaté particulièrement pour le domaine des technologies mobiles, dans lequel les abréviations très importantes pour le domaine sont rarement utilisées en forme développée (*IP, GSM, WLAN*, etc.), et on a fini par les inclure dans les RTLs.

RTLs bilingues. La construction des RTLs bilingues a confirmé que le lien d'équivalence entre les TS et les TC n'est pas toujours univoque. Un TS peut avoir plusieurs équivalents en langue cible, et vice versa. Les raisons sont les suivantes :

1. Synonymie dans la langue cible (EN *wind tunnel* = FR *soufflerie* = FR *tunnel aérodynamique*, EN *airfoil* = FR *surface portante* = FR *profil d'aile*) ;
2. Synonymie dans la langue source (EN *wind speed* = FR *vitesse du vent* et EN *wind velocity* = FR *vitesse du vent*) ;
3. Homonymie dans une des langues (EN *wind turbine* = FR (1) *éolienne*, (2) *turbine éolienne*).

Il a été convenu que les RTLs bilingues allaient contenir les alignements « 1-1 », alors que les termes synonymiques seraient inclus comme variantes dans chaque langue. Pour choisir l'entrée principale pour un terme, nous avons pris en compte sa fréquence dans le corpus du domaine : ainsi, pour EN *wind tunnel*, nous avons choisi comme terme équivalent FR *soufflerie* (fréquence de 64), tandis que FR *tunnel aérodynamique* (fréquence de 1) a été marqué comme sa variante synonymique.

Nous constatons que la construction des RTLs comporte un volet d'expertise à la fois terminologique et linguistique. Certains critères peuvent être formulés pour aider un terminologue à sélectionner les termes caractéristiques d'un domaine donné, mais la constitution de liste terminologique de référence parfaite s'avère un idéal.

4.3 Résultats de l'évaluation

Les RTLs construites ont été utilisées dans l'évaluation des outils de traitement des termes, développés dans le cadre du projet TTC. Nous rapportons ici les résultats obtenus avec un des outils, TermSuite, pour illustrer ce que l'extraction automatique, dans l'état actuel, peut apporter. Nous analysons des erreurs produites par l'extracteur, et nous les comparons avec les difficultés qu'un humain rencontre au cours de la construction d'une liste de termes de référence.

TermSuite [8] effectue l'extraction des termes, aussi bien mono- que polylexicaux, à partir de corpus spécialisés. Les termes monolexicaux sont extraits en s'appuyant sur leur fréquence et leur catégorie grammaticale, les termes polylexicaux sont extraits à l'aide des patrons morphosyntaxiques définis pour chaque langue. Des termes extraits des corpus comparables sont ensuite alignés, c'est-à-dire que les liens d'équivalence (traduction) sont établis. L'alignement des termes monolexicaux s'effectue avec la méthode distributionnelle (Rapp, 1999), l'alignement des termes polylexicaux se fait avec la méthode compositionnelle (Grefenstette, 1999) et la méthode mixte (compositionnelle étendue avec la construction des vecteurs de contextes) (Morin et Daille, 2012).

Nous rapportons ici les résultats le l'extraction monolingue (EN, FR) et de l'alignement bilingue des termes anglais-français.

8. Pour les expériences rapportées, la version TermSuite 1.4 a été utilisée.

4.3.1 Extraction monolingue

Il y a plusieurs manières de comparer la liste de candidats avec la liste de référence. Dans Mustafa El Hadi et al. (2006) cinq critères sont proposés :

— un candidat correspond exactement à une entrée du référentiel ;
— un candidat n'est pas dans le référentiel, mais il est admis comme terme par un expert ;
— pour un candidat polylexical, au moins deux composants du candidat sont présents dans le référentiel ;
— un composant du candidat fixé (par exemple, la tête du syntagme) est présent dans le référentiel ;
— un des composants, indifféremment, du candidat est présent dans le référentiel ;

Dans l'évaluation effectuée, seul le premier critère a été utilisé : un candidat doit correspondre exactement à un terme ou à une variante présents dans la RTL.

La précision, le rappel et la F-mesure ont été calculés. La précision est égale à la taille de l'intersection entre les termes candidats extraits (CT) et les termes de référence (RT), divisée par le nombre de termes candidats :

$$P = \frac{|CT \cap RT|}{|CT|} \tag{4.1}$$

Le rappel est égal à la taille de l'intersection entre les candidats termes extraits (CT) et les termes de référence (RT), divisée par le nombre de termes dans la RTL :

$$R = \frac{|CT \cap RT|}{|RT|} \tag{4.2}$$

La F-mesure dite « balancée » (Virpioja et al., 2011) (avec les poids équivalents du rappel et de la précision) est calculée comme la moyenne harmonique entre les deux mesures :

$$F = \frac{2 \times P \times R}{P + R} \tag{4.3}$$

La figure 4.1 présente la F-mesure obtenue en fonction du nombre de candidats termes monolingues extraits (CT), triés par la valeur de spécificité [9].

Le problème de l'utilisation de ces mesures pour l'extraction terminologique monolingue est qu'il est impossible de sélectionner d'avance tous les termes d'un corpus, ni de définir leur nombre (Vintar, 2010). Sachant que l'on compare une très longue liste de candidats (plusieurs dizaines de milliers de candidats) avec une courte liste de références (entre 130 et 200 termes en incluant les variantes), la précision obtenue s'avère très basse. Cela affecte la F-mesure qui ne dépasse pas 37 % pour EN et 23 % pour FR pour le domaine des technologies mobiles, et 15 % pour EN et 18 % pour FR pour le domaine de l'énergie éolienne). Certains CT corrects ne sont pas présents dans les RTLs, ils sont donc comptés comme incorrects avec ce type d'évaluation. Une expertise humaine, au moins pour les premiers candidats de la liste, est donc souhaitable même dans le cas de l'utilisation d'une liste de référence pour l'évaluation.

Néanmoins, le nombre de CT incorrects est aussi élevé. Après avoir examiné la sortie de l'outil, nous énumérons les erreurs et les difficultés les plus courantes :

1. Des candidats extraits sont mal formés (le syntagme mal découpé) : FR **ferme éolien de puissance* [10], le terme correct serait (dans sa forme lemmatisée) *ferme éolien*, ou éventuellement (avec un statut terminologique moins certain) *ferme éolien de puissance moyen (ferme éolienne de puissance moyenne)*.

9. Les résultats sont tirés du Rapport T3-4 du projet TTC.
10. Tous les constituants des termes polylexicaux en sortie de TermSuite sont lemmatisés, la forme *ferme éolien* est donc correcte.

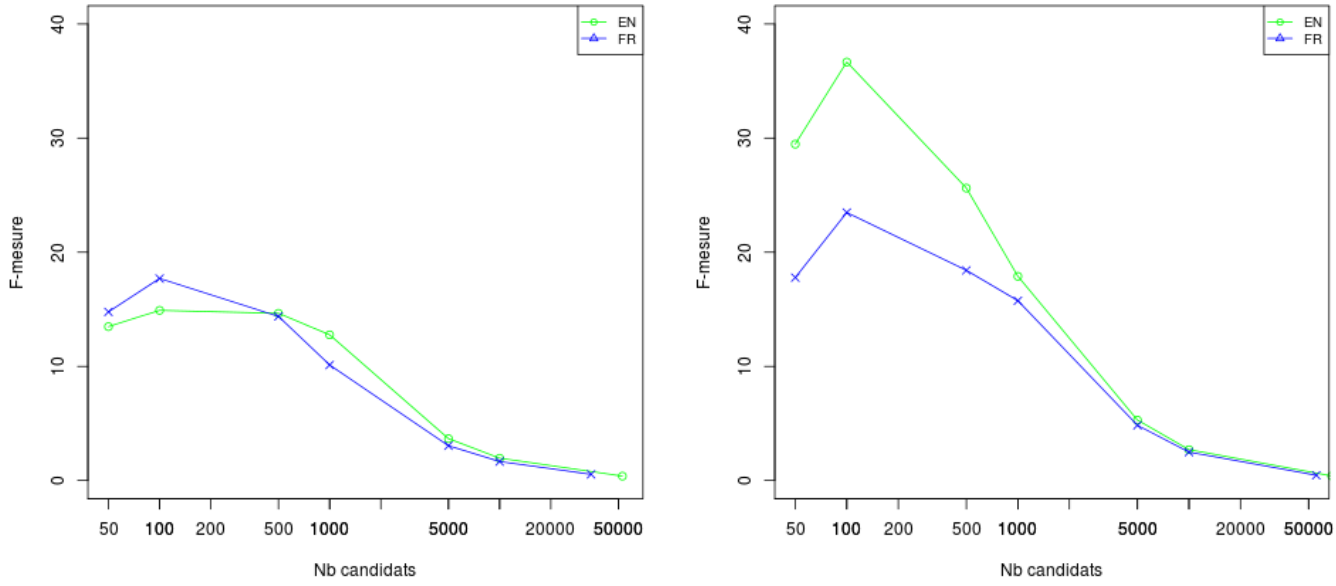

FIGURE 4.1 – Extraction monolingue : F-mesure en fonction du nombre de candidats termes extraits pour le domaine de l'*énergie éolienne* (à gauche) et des *technologies mobiles* (à droite)

2. Parmi les candidats extraits, certains n'ont pas de haut degré de spécialité (EN *find* 'trouver' dans le corpus du cancer du sein) ou appartiennent à des domaines adjacents (EN *gearbox* 'boîte de vitesses' dans le corpus de l'énergie éolienne).
3. Parmi les CT polylexicaux, certains représentent des collocations d'un terme monolexical, mais ne sont pas des termes : EN *hub height* 'hauteur du moyeu' est une collocation du terme *hub* 'moyeu'.

Les deux derniers points sont, néanmoins, discutables : il faudrait une étude approfondie des contextes pour chaque unité lexicale afin de prouver ou contester son statut terminologique dans le domaine. Nous avons d'ailleurs vu, en construisant les RTLs, que de tels cas sont également problématiques pour un humain qui décrit la terminologie d'un domaine. Le problème des candidats mal formés est, par contre, spécifique à l'extraction automatique.

4.3.2 Extraction bilingue

Pour évaluer l'alignement bilingue, les termes sources (TS) appartenant à la RTL bilingue ont été alignés à l'aide de TermSuite avec les termes cibles (TC) extraits au niveau monolingue. Un alignement est considéré comme correct si le TS correspond exactement au TC de la RTL bilingue ou à une de ses variantes. Pour chaque TS, un ou plusieurs alignements sont proposés par TermSuite. Nous fixons le nombre maximal d'alignements candidats à prendre en compte (Top N).

La précision de l'alignement pour le Top N est calculée comme le nombre de TS pour lesquels au moins un alignement correct a été proposé parmi N candidats, divisé par le nombre de TS alignés :

$$\text{Précision} = \frac{\text{nb. TS alignés correctement}}{\text{nb. TS alignés}} \tag{4.4}$$

Le rappel de l'alignement pour le Top N est calculé comme le nombre de TS pour lesquels au moins un alignement correct a été proposé parmi N candidats, divisé par le nombre total de TS dans RTL :

$$\text{Rappel} = \frac{\text{nb. TS alignés correctement}}{\text{nb. total TS dans RTL}} \tag{4.5}$$

TABLE 4.5 – Évaluation d'extraction terminologique bilingue avec TermSuite en prenant en compte le Top 100 des traductions candidates pour chaque terme source.

	Nb termes		Précision (%)		Rappel (%)		F-mesure (%)	
	mobiles	éoliennes	mobiles	éoliennes	mobiles	éoliennes	mobiles	éoliennes
SWT	58	46	60	67	26	39	36	49
MWT	42	57	100	91	45	40	62	56

Nous rapportons ci-dessous les résultats pour les termes monolexicaux (SWT) et polylexicaux (MWT) séparément, obtenus en fixant le nombre maximal d'alignements candidats pour un TS à 100 (cf. tableau 4.5). Les résultats sont tirés du Livrable 4.4 du projet TTC.

Nous constatons qu'une haute précision est obtenue pour les termes polylexicaux : 100 % (technologies mobiles) et 91 % (énergie éolienne). Par contre, le rappel pour les termes polylexicaux ne dépasse pas 45 % (technologies mobiles). La F-mesure est de 62 % pour les technologies mobiles et de 56 % pour l'énergie éolienne. Pour les termes monolexicaux, la précision et le rappel sont inférieurs : la précision se situe entre 60 % et 67 %, le rappel entre 26 % et 39 %, la F-mesure entre 36 % et 49 %.

Les erreurs d'alignement les plus courantes sont les suivantes :

1. Des termes ne sont pas alignés parce que leurs traductions ne sont pas présentes dans le corpus cible (ou présentes avec une fréquence peu élevée), nos corpus étant d'une taille modeste et comparables entre eux (et non parallèles) : le terme EN *domestic wind turbine* n'est pas aligné car une traduction potentielle FR *éolienne domestique* apparaît dans le corpus avec une fréquence de 4, et une autre traduction possible *turbine éolienne domestique* n'y apparaît pas.
2. Des TS monolexicaux complexes qui devraient être alignés avec des TC polylexicaux, ne le sont pas (en l'état, TermSuite ne prévoit pas ce type d'alignement) : EN *airfoil* 'profil aérodynamique' est aligné par la méthode distributionnelle avec les candidats monolexicaux *pale, éolienne, profil, aérodynamique*, etc., mais non avec *profil aérodynamique* en tant qu'unité.

Le deuxième type d'erreur explique pourquoi l'alignement des termes polylexicaux anglais (généralement traduits par des termes polylexicaux français) est plus précis que l'alignement des termes monolexicaux.

4.4 Bilan

Dans cette partie, nous avons examiné les principales méthodes d'extraction terminologiques (monolingue et bilingue) à partir des corpus spécialisés. Nous avons également parcouru les techniques utilisées pour évaluer et comparer les différentes méthodes d'extraction.

Nous avons participé à l'évaluation d'extracteurs automatiques des termes. La tâche à effectuer était la construction des listes de référence des termes de domaines de spécialité, de manière semi-manuelle. Nous avons discuté des aspects techniques de l'élaboration de ces listes et des difficultés rencontrées. Nous avons rapporté les résultats de l'évaluation de l'extraction monolingue (EN, FR) et bilingue (EN-FR), effectuée par le logiciel TermSuite, l'évaluation étant menée en utilisant les RTLs construites.

L'évaluation mise en place avec une liste de référence a ses défauts. Notamment, des candidats extraits proprement ne sont pas comptés comme corrects s'ils ne sont pas inclus dans la liste, et le choix des termes dans la liste influence les résultats. Dans le même temps, la liste de référence est construite par un humain, et de ce fait elle peut difficilement être exhaustive. En revanche, une telle liste peut être facilement réutilisée pour d'autres évaluations, et permet de comparer l'efficacité de plusieurs systèmes.

Enfin, nous avons comparé la construction supervisée des RTLs avec l'extraction entièrement automatique (sur l'exemple des candidats termes extraits et alignés par TermSuite). Dans les deux cas, nous devons

faire face au problème de la gradualité du degré de spécificité de l'unité lexicale. Parmi les problèmes spécifiques à la méthode automatique, nous pouvons mentionner le découpage des termes polylexicaux qui n'est pas toujours bien accompli, la sensibilité de la méthode à la taille des corpus utilisés et à leur degré de comparabilité (pour l'extraction multilingue), et l'établissement d'équivalence échoué entre les termes monolexicaux complexes de la langue source et leurs traductions polylexicales de la langue cible. Le dernier point a inspiré la suite de nos travaux.

Segmentation des termes composés

Les problèmes rencontrés au cours de l'évaluation de l'extraction terminologique à partir de corpus comparables confirment notre hypothèse que les termes composés (i.e. à la fois monolexicaux et sémantiquement complexes) requirent un traitement particulier dans un contexte multilingue. La solution qui s'impose est de les segmenter afin d'utiliser leurs composants pour plusieurs tâches.

Ainsi, pendant l'extraction bilingue, la segmentation des termes composés aide à les aligner avec des termes polylexicaux d'une autre langue (Weller et Heid, 2012), ce qui n'est pas possible avec une méthode distributionnelle d'alignement. De même, au niveau monolingue, les termes complexes ont souvent des variantes polylexicales (terme *insulino-dépendant* - variante *dépendant de l'insuline*), et pour les regrouper, la segmentation est également utile.

Le phénomène de la composition est productif dans de nombreuses langues appartenant à des familles différentes (germaniques, ouraliennes, slaves, etc.). Pour donner un ordre de grandeur, Alfonseca et al. (2008) démontrent que dans un corpus d'actualités de langue allemande, entre 5-7 % de lexèmes et 43-47 % d'occurrences sont des mots composés. Cependant, les applications automatiques ont tendance à les traiter comme des mots simples du fait qu'ils comportent un seul mot graphique. Pour différentes tâches du TAL, il est utile d'identifier les composés dans les textes et d'appliquer une méthode de segmentation. Il a été demontré que la reconnaissance et la segmentation des composés sont bénéfiques pour la traduction automatique (Macherey et al., 2011), la recherche d'information monolingue (Braschler et Ripplinger, 2004) et multilingue (Chen et Gey, 2001), la reconnaissance automatique de la parole (Larson et al., 2000), etc. Notre travail s'inscrit dans le traitement des domaines de spécialité pour lesquels la composition est encore plus fréquente que dans la langue générale. La segmentation des termes composés s'avère donc une étape nécessaire.

Dans cette partie nous nous penchons sur le phénomène de la composition et les méthodes existantes de segmentation automatique avant de tenter de proposer une nouvelle approche adaptée pour les langues spécialisées. Notre méthode est multilingue, à la fois fondée sur l'utilisation des corpus, adaptable à la langue et au domaine de spécialité traités, et permettant d'intégrer des ressources lexicales et des connaissances linguistiques. L'objectif de notre méthode est d'identifier les termes composés et de les segmenter, ce qui signifie non seulement repérer les frontières entre les composants graphiques, mais aussi mettre en correspondance les composés et les unités lexicales à partir desquelles ils sont formés. Nous présentons les expériences effectuées sur quatre langues : l'allemand, le russe, l'anglais et le français ; et sur deux domaines de spécialité : *énergie éolienne* et *cancer du sein*. Enfin, nous comparons les résultats obtenus à ceux d'autres méthodes de l'état de l'art : une méthode probabiliste, une méthode fondée sur le corpus, et une méthode lexicale.

5

Composition morphologique

Dans ce chapitre, nous étudions le phénomène de la composition sous l'angle de ses différentes interprétations dans la littérature (section 5.1). Nous nous limitons à la composition dite morphologique. Sur une base à la fois théorique et applicative, nous dressons une typologie des phénomènes que nous allons traiter dans ce travail. Dans la section 5.2, nous résumons les difficultés liées au traitement des composés en TAL et nous faisons un tour d'horizon des méthodes proposées pour l'identification et la segmentation automatique des composés, ainsi que les moyens d'évaluer la qualité de segmentation.

5.1 Spécifications linguistiques des mots composés

5.1.1 Définitions

Composition D'après (Lehmann et Martin-Berthet, 2008, p. 217), « la composition consiste à former un mot en assemblant deux ou plusieurs mots ». Cette définition donne rapidement une idée de la composition, elle est cependant insuffisante car elle n'explique pas comment on identifie un « mot », une notion cruciale en linguistique qui a elle-même plusieurs définitions. Dans (Bussmann, 1996), la composition est définie comme « la combinaison de deux ou plusieurs morphèmes ou séries de morphèmes (= mots), libres dans d'autres contextes, qui forment un composé »[1]. Cependant, les composants d'un composé, tout en étant sémantiquement indépendants, peuvent avoir une forme différente des unités lexicales autonomes, par exemple FR *hormonosensible* = *hormone* + *sensible*.

Voilà pourquoi Bauer (1983) définit le mot composé comme un lexème contenant « deux ou plusieurs bases potentielles ».[2] Également d'après (Gardes-Tamine, 2010, p.88), la composition est « la juxtaposition de deux éléments qui peuvent servir de base à des dérivés ». Le résultat du processus de composition est un composé.

La délimitation des phénomènes couverts par le terme « composition » dépend de l'approche adoptée et de la tradition linguistique. Ainsi, dans la grammaire traditionnelle allemande, un mot composé, « *Kompositum* », correspond forcément à une unité graphique, de même que dans la grammaire traditionnelle russe

1. « Combining of two or more otherwise free morphemes or series of morphemes (= words) to form a compound » (Bussmann, 1996).

2. « A compound lexeme (or simply a compound) can thus be defined as a lexeme containing two or more potential stems ». (Bauer, 1983, p.28)

(cf. (Plungian, 2000). Par contre dans les grammaires anglaise (par exemple (Bauer, 1983)) et française, la composition s'étend sur « l'ensemble des unités lexicales composées de plusieurs mots » (Lehmann et Martin-Berthet, 2008, p. 222), y compris de plusieurs mots graphiques. Nous allons suivre cette conception large de la composition. Par conséquent, nous pouvons dire que tous les termes complexes tels que nous les avons définis dans la section 2.1 sont des composés. Nous allons donc avoir besoin d'introduire d'autres dénominations pour distinguer les composés monolexicaux (contenant un mot graphique) des composés polylexicaux (ceux en contenant plusieurs). Nous avons choisi les termes « composés morphologiques » et « composés syntagmatiques ».

Composés morphologiques vs. syntagmatiques Les composés syntagmatiques (Cusin-Berche, 2003) sont des composés dont les éléments sont séparés par un espace, tout en formant une unité de sens, e.g. FR *barre d'outils*, contrairement aux composés morphologiques qui correspondent à un seul mot graphique, e.g. EN *toolbar*.

5.1.2 Quelques précisions sur le choix des termes

Les linguistes s'intéressent aux groupes lexicaux supérieurs à un mot graphique depuis les années 1960, suite à l'apparition de la notion de « syntagme » chez Saussure et grâce aux premiers travaux en traduction automatique (Léon, 2004). Ainsi, Charles Bally a défini un composé comme « un syntagme virtuel caractérisé qui désigne, en la motivant, une idée unique » (Bally, 1965, p.94).

Pottier (1962) a introduit la notion de « lexie », unité de sens, de segmentation et de traduction. Il distingue les lexies simples (*chaise*), composées (*cheval-vapeur*) ou complexes (*pomme de terre, cheval de frise*). Les lexies simples et composées sont identifiées par un programme informatique comme une séquence de lettres entre deux séparateurs de mots.

Les lexies complexes cependant ne sont pas triviales à identifier. Pour cela, en plus du critère d'unicité de l'objet désigné, Pottier propose des critères syntaxiques telles que la non-séparabilité (*pomme de terre* - **pomme ronde de terre*), la non-qualification (*prendre la mouche* - **prendre la grosse mouche*) et la non-quantification (*plaque minéralogique* - **plaque peu minéralogique*), la non-prédicativité (*plaque minéralogique* - **cette plaque est minéralogique*). Il reste toutefois toujours des cas peu évidents, comme par exemple la lexie *crise de croissance* pour laquelle le critère de la non-prédicativité fonctionne (on ne peut pas dire **cette crise est de croissance*), mais pas celui de la non-qualification (*crise aiguë de croissance*).

Benveniste (1974) a introduit un type particulier de composition, qui mérite pour lui d'être qualifié de phénomène indépendant, la « synapsie ». La synapsie est « un groupe entier de lexèmes, reliés par divers procédés, et formant une désignation constante et spécifique » (Benveniste, 1974, p.172) : *robe de chambre, salle à manger, gardien d'asile de nuit*, etc. Il énumère les principales caractéristiques de synapsie pour la langue française : « 1) la nature syntaxique (non morphologique) de la liaison entre les membres ; - 2) l'emploi de joncteurs à cet effet, notamment *de* et *à* ; - 3) l'ordre déterminé + déterminant des membres ; - 4) leur forme lexicale pleine, et le choix libre de tout substantif ou adjectif ; - 5) l'absence d'article devant le déterminant ; - 6) la possibilité d'expansion pour l'un ou pour l'autre membre ; - 7) le caractère unique et constant du signifié » (ibid.).

Les synapsies sont extrêmement productives dans les langues de spécialité grâce à cette capacité d'expansion des membres qui spécifie le désigné : *volet de courbure, volet à fente* → *volet de courbure à fente* (domaine d'aviation). (Benveniste, 1974, p.174)) :

> « la synapsie prodigue sans trêve ses créations. Tous les vocabulaires techniques y font appel, et d'autant plus aisément qu'elle seule permet la spécification du désigné, et la classification des séries par leur trait distinctif. Son extrême flexibilité paradigmatique fait de la synapsie l'instrument par excellence des nomenclatures ».

Revenons au premier critère de synapsie pour Benveniste : la relation syntaxique entre les membres. La nature des liens entre les éléments est parfois vue comme le trait distinctif entre les composés syntagmatiques et les composés morphologiques. Ces derniers tiennent d'ailleurs leur appellation du fait qu'ils ont initialement été traités dans le cadre de la morphologie. Benveniste (1974) met en évidence la nature syntaxique des liens à l'intérieur des composés, y compris des composés morphologiques. Pour lui, dans la composition, la construction syntaxique est toujours derrière la structure formelle. Par exemple, *oiseau-mouche* a la structure « qui est », « oiseau qui est une mouche » (qui ressemble à une mouche) ; EN *blue-eyed* a la structure « être à », « quelqu'un aux yeux bleus », etc. Au composé, Benveniste oppose le *conglomeré*, « construction complexe qui se soude en un bloc, sans que les éléments soient mutilés ou altérés » (Benveniste, 1974, p.171) : *va-nu-pied, justaucorps, désormais*, etc. Dans un conglomeré, on devine plus ou moins les éléments (cela dépend de l'âge du lexème), mais sans avoir l'impression d'un ensemble composé, « le sentiment de la composition est déjà aboli ». Les éléments n'ont plus de rapport syntaxique, ils sont complètement soudés.

Dans les recherches plus récentes, il n'y a toujours pas d'accord général sur la nature morphologique ou syntaxique des composés (cf. Ackema et Neeleman (2010) sur la compétition entre les modules morphologique et syntaxique). D'un côté, on ne peut pas ignorer la « syntaxe interne » (Aronoff, 1993) présente dans ce type de constructions. Des transformations syntaxiques peuvent également s'appliquer aux composés. De l'autre côté, la composition se rapproche de la dérivation sur certains points tels que l'allomorphie ou les patrons accentuels différents de ceux des phrases (Scalise et Vogel, 2010). Certains composés n'obéissent pas aux règles syntaxiques de la langue (FR *porte-monnaie* : l'absence d'article (Lehmann et Martin-Berthet, 2008)). D'autres formations obéissent à ces règles (FR *coffre-fort, trou noir*) et de ce fait peuvent être considérées comme des « expressions construites par la syntaxe » et non comme des composés (Fradin, 2003).

Nous partageons l'idée que la nature des liens entre les éléments d'un composé n'est pas purement morphologique, même dans le cas d'un composé contenant un seul mot graphique. Néanmoins, de manière conventionnelle, nous conserverons dans cette thèse le terme « composé morphologique » par opposition au « composé syntagmatique ».

D'une part, le fait de réunir les deux types sous un terme général « composés » reflète l'idée, très importante en TAL, que les unités mono- et polylexicales sont des unités de même niveau. Ainsi, dans des langues différentes, le même concept peut être incarné par des unités de types différents, e.g. la majorité des composés morphologiques allemands se traduisent en français par des composés syntagmatiques (DE *Rotorblatt* - FR *pale de rotor*). La frontière entre les composés dits morphologiques et syntagmatiques est d'autant plus floue que la forme graphique peut varier (Neveu, 2004, p. 89) :

> « Dans la mesure où elle peut connaître des variations dans un même état de langue, la réalisation graphique d'un mot composé est assez aléatoire (ex. *contre-pied / contrepied*, *mot-clé / mot clé*, etc.) ».

D'autre part, la distinction formelle qui est derrière les deux dénominations est également très importante en TAL, notamment parce que les composés morphologiques et les composés syntagmatiques induisent des difficultés différentes pour les systèmes de traitement automatique du texte : les premiers nécessitent d'être segmentés en éléments porteurs de sens, tandis que les derniers doivent au contraire être identifiés comme une unité de sens. Dans cette partie nous nous focaliserons sur la problématique de la reconnaissance et de la segmentation des composés morphologiques (ultérieurement « composés » tout court).

5.1.3 Classifications des composés

De nombreuses classifications des composés ont été proposées dans la littérature. La première sorte de classification est basée sur l'interprétation sémantique. D'après cette classification (Bussmann, 1996), on distingue les composés « déterminants » (« *determinative compounds* »), « possessifs » (« *possessive compounds* ») et « copulatifs » (« *copulative compounds* »). Dans les premiers, un élément est sémantiquement déterminé par l'autre, il vient spécifier l'autre, par exemple DE *Wandschrank* (lit. 'mur + placard') 'placard mural'. Dans le deuxième type, aussi connu sous le terme sanskrit « *bahuvrihi* », la relation est possessive. On peut paraphraser « celui qui a » : EN *redhead* (lit. 'roux + tête') 'roux, rouquin' = « celui qui a une tête (chevelure) rousse ». Dans le troisième type (sanskrit « *dvandva* »), les deux éléments sont sémantiquement équivalents et, mis ensemble, ils réfèrent à un nouveau concept, e.g. EN *author-editor*, DE *Hosenrock* (lit. pantalon + jupe) 'jupe-culotte'. Bauer (1983) appelle le dernier type « composés apositionnels » (« *appositional compounds* »), et il réserve le terme « composés copulatifs » seulement aux composés dont les éléments nomment des entités séparées, e.g. *Alsace-Lorraine*.

Une autre sorte de classification est basée sur les relations grammaticales entre les composants, comparables aux relations syntaxiques. Nous citons ici un exemple d'une telle classification tiré de (Scalise et Vogel, 2010). Cette classification distingue les composés « subordonnés » (« *subordinate compounds* ») e.g. EN *taxi driver*, « attributifs » (« *attributive compounds* ») e.g. EN *ice cold* et « coordonnés » (« *coordinate compounds* ») e.g. EN *poet painter*.

On pourrait établir certaines correspondances entre les deux classifications, mais elles ne sont pas équivalentes. Ainsi, on peut mettre en correspondance les composés copulatifs et coordonnés, les composés possessifs avec les attributifs, en revanche tous les attributifs ne sont pas possessifs (cf. EN *ice cold*). Les éléments des composés déterminants sont parfois liés par une relation de subordination (EN *taxi driver*), et parfois par une relation d'attribution (EN *living room*).

On distingue également les composés « endocentriques » des composés « exocentriques ». Les premiers peuvent être présentés comme hyponymes de leur tête : EN *taxi driver* est un type de conducteurs, EN *blackbird* est un type d'oiseaux, etc. Les composés exocentriques s'appellent ainsi parce que leur « centre sémantique » n'est pas égal à un des éléments et se trouve « à l'extérieur » du composé : *pickpocket* (lit. 'voler + poche') désigne une personne qui vole, et non la poche, EN *redhead* désigne une personne qui a des cheveux roux, et non seulement sa tête, *Alsace-Lorraine* désigne le territoire formée par Alsace et Lorraine. Il existe un autre type particulier de composés dits « néoclassiques », c'est-à-dire des composés ayant un ou plusieurs éléments d'origine latine ou grecque (Namer, 2009) : *hydrologie, cardiovasculaire, micro-ordinateur*, etc. Ils sont également appelés « composés savants » car ce type de formation est très productif dans les textes scientifiques. La particularité de cette formation est que les éléments néoclassiques (*cardio-, hydro-, -logie, micro-*) n'apparaissent jamais indépendamment dans les textes, mais seulement en combinaison avec d'autres éléments (Amiot et Dal, 2008; Namer, 2009). En même temps, les éléments néoclassiques ne peuvent pas être assimilés aux affixes parce que, contrairement à eux, 1) ils sont porteurs de sens et peuvent former des mots sans autre racine (*hydrologie*) ; 2) certains d'entre eux peuvent apparaître en position initiale aussi bien que finale : *patho-* dans *pathologie* et *-pathie* dans *cardiopathie*, les deux formes provenant de la racine grecque *pathos* (Bauer, 1983). Les composés néoclassiques sont parfois vus comme une zone de frontière entre la dérivation et la composition (Béchade, 1992). Le terme « composition populaire » est employé par opposition à la « composition savante » (Lehmann et Martin-Berthet, 2008). Dans ce travail, nous utilisons les termes « composés néoclassiques » et « composés natifs ».

5.1.4 Cas périphériques

En dehors des composés néoclassiques, il existe d'autres cas à la frontière entre dérivation et composition. Il est parfois difficile de définir si un composant est indépendant ou non à cause du processus de grammaticalisation constamment présent dans les langues. Par exemple, le suffixe anglais *ful* a été formé à partir du mot *full*, 'plein'. Actuellement il est grammaticalisé et ne peut pas être analysé comme un composant indépendant. Tant que le processus n'est pas abouti, en synchronie il est parfois difficile de trancher entre une racine et un suffixe. Ainsi, le mot EN *work* participe en tant que racine dans les composés *workflow*, *artwork*, mais dans les mots *wirework*, *woolwork* il peut être qualifié de suffixe avec la signification de « décoration », « ornementation » (Dictionnaire Collins). Un segment du mot peut aussi se transformer en une base ou un suffixe productif de composition : « info(rmatique) > infographiste, inforoute, infogérance ; vélo(cipède) > vélodrome, vélomoteur, véloski ; (informa)tique > domotique, bureautique, robotique, monétique ; logi(ciel) > ludiciel, didacticiel, courriel » (Cusin-Berche, 2003, p.99).

Pour les mêmes raisons, il n'est pas toujours évident de fixer la limite entre la composition et la préfixation. Par exemple, les adverbes FR *bien* et *mal* participent à la formation des séries de mots (*bienfait, bien-être, maladroit, malvoyant*, etc.). Du fait de cette productivité, les formations du type *malvoyant* sont qualifiées de préfixées, contrairement à *clairvoyant* (composé). Les formations en *outre-* ou en *plus-*, elles, sont moins productives et se situent entre les deux pôles. Cela permet de parler du continuum plutôt que de deux catégories opposées (Lehmann et Martin-Berthet, 2008). Certains préfixes d'origine classique sont très proches des racines néoclassiques, cf. le préfixe *bi-* vs. la racine néoclassique *uni-* (selon Béchade (1992)). Pour les distinguer, on peut s'appuyer sur leur origine (les préfixes classiques remontent aux mots grammaticaux tandis que les racines néoclassiques proviennent des mots pleins) et sur leur âge : les préfixes sont entrés dans l'usage beaucoup plus tôt (Béchade, 1992). Lehmann et Martin-Berthet (2008) qualifient *télé* comme la racine néoclassique dans *téléphone* et comme préfixe dans *télécharger, télésurveillance*. La préfixation se rapproche d'autant plus de la composition que les préfixes sont souvent porteurs d'un composant sémantique : *prévoir* = « voir avant », *antichar* = « contre les chars », etc. « La préfixation apparaît ainsi comme la condensation lexicale d'un syntagme, de même que la composition » (Lehmann et Martin-Berthet, 2008, p.168).

D'autres formations contenant plus d'une racine peuvent aussi contenir un affixe dérivationnel :

EN blue-eyed : $[[\text{blue} + \text{eye}]_{NP} + \text{ed}]_{ADJ}$;
RU голубоглазый 'aux yeux bleus' : $[[\text{golub 'bleu'} + \text{o} + \text{glaz 'œil'}]_{NP} + \text{yj}]_{ADJ}$ [3]

Bauer (1983) n'inclut pas ce phénomène dans la composition. Pour de tels cas il emploie le terme plus général de « mots complexes » (« *complex words* »). Certains chercheurs les qualifient comme un type particulier de composition, les « composés parasynthétiques » (Melloni et Bisetto, 2010).

Cusin-Berche introduit la notion de *compocation*, l'amalgame de la composition et de la troncation, par exemple *héliport (héli(coptère aéro)port)*, *imprimaticien (imprim(eur inform)aticien), micromiser (micro (mini)miser)*. Ce procédé, comme celui de la formation des mots-valises, acronymes et sigles « favorisent la réalisation d'unités qui sont moins immédiatement décomposables et interprétables, et qui, de ce fait, impliquent un détour de type étymologique » (Cusin-Berche, 2003, p.34). Si c'est le cas pour un humain, l'analyse automatique de telles formations s'avère difficilement réalisable.

5.1.5 Délimitation applicative et typologie des phénomènes traités

Nous nous plaçons dans un cadre multilingue et opérationnel, notre objectif étant l'identification et la segmentation des termes composés reliés aux domaines spécialisés. Voilà pourquoi les classifications des composés basées sur les relations sémantiques entre les composants et la tête de composé, matérialisée ou non par un composant, ne semblent pas adaptées à notre travail. Quant à la classification du point de vue

3. Dans tout le document, les mots russes sont translittérés selon le système de translittération ISO 9.

des relations grammaticales entre les composants, elle peut avoir son utilité pour le traitement automatique des composés. Cependant l'identification automatique des relations syntaxiques à l'intérieur d'un syntagme est en soi une tâche non triviale, elle l'est d'autant plus à l'intérieur d'un composé morphologique, faute de frontières explicites entre les composants. Nous préférons adopter une classification opérationnelle pour la tâche de segmentation qui s'appuie sur les ressources lexicales.

Pour établir une telle classification, nous examinons les composants formant un composé, et nous définissons quatre caractéristiques binaires d'un composant :

1. Autonomie du composant (c'est-à-dire, apparaît-il indépendamment dans les textes de cette langue comme un mot simple ou non) ;
2. Recensement dans un lexique (le composant ou sa forme canonique, si le composant n'est pas autonome, appartiennent-ils à un dictionnaire général de cette langue ?) ;
3. Attestation en corpus (le composant ou sa forme canonique sont-ils attestés dans un corpus de textes de cette langue ?) ;
4. Caractère multilingue du composant (le composant a-t-il des équivalents phonétiquement très proches dans d'autres langues, y compris des langues appartenant à des familles différentes ?) ;

Par rapport à ces caractéristiques, on peut identifier des sous-ensembles ou des types de phénomènes. Ainsi se distinguent les composés **néoclassiques** et **natifs**. Les composants néoclassiques ne sont pas autonomes, ne sont pas attestés ni dans les corpus, ni dans les dictionnaires généraux (même si des exceptions existent), et par conséquent ils nécessitent un traitement particulier ou une ressource lexicale supplémentaire. Dans le même temps, les éléments néoclassiques se ressemblent dans différentes langues, étant donné leur origine commune. Les composants des composés natifs peuvent être autonomes ou non. Contrairement aux composés néoclassiques, leur forme canonique appartient à un dictionnaire et apparaît dans les textes, mais ils ne sont pas multilingues.

En utilisant ces caractéristiques, un autre type de composés émerge : les **composés empruntés**, c'est-à-dire des composés dont un des composants est emprunté à une autre langue, mais pas au grec ou au latin (généralement il s'agit de l'anglais) ; e.g. RU web-страница 'page web' ou DE *Repowering-Leitfaden* 'repowering + consignes'. Les composants empruntés sont parfois adaptés phonétiquement ou graphiquement à la langue cible. Par exemple, RU прайс-лист *prajs-list* 'price-list', où прайс est le mot EN *price* translittéré en alphabet cyrillique. Les composants empruntés ne sont pas forcément multilingues (un emprunt singulier peut avoir lieu, comme DE *Repowering-Leitfaden*). Comme les composants néoclassiques, ils ne font pas partie d'un dictionnaire général, par contre, ils peuvent apparaître dans un corpus. Enfin, ils sont autonomes et, contrairement aux néoclassiques, sont facilement séparables, étant souvent attachés à la seconde base par un trait d'union.

Nous avons rajouté deux autres catégories qui sont frontalières : l'une qui se situe entre la composition et la dérivation ; l'autre, entre la composition et la syntaxe.

Les mots **préfixés** tels que *bipale* ou *sous-système*, même si ce ne sont pas des composés au sens strict, peuvent être traités dans le même cadre applicatif que les composés, parce que (1) les préfixes sont facilement séparables de la base, et (2) ces mots ont souvent une structure équivalente dans d'autres langues. Du point de vue des caractéristiques énumérées, les préfixes ne sont pas autonomes, pas forcément multilingues, n'appartiennent pas aux dictionnaires et ne sont pas attestés dans le corpus sans autres éléments.

La dernière catégorie se trouve aux frontières entre les composés dit morphologiques et les constructions purement syntaxiques. Cette catégorie comprend les composés appelés « déverbaux » dans la tradition anglophone (« *verbal compounds* » (Roeper et Siegel, 1978)), par exemple EN *anthracycline-based* qui peut être paraphrasé comme « basé(e) sur l'anthracycline ». La signification de l'ensemble est parfaitement claire à partir des composants. Un autre cas est le composé français *fréquence-puissance* qui est, au contraire, non compréhensible sans contexte, et doit être considéré dans l'expression *réglage fréquence-puissance*. Ici, les deux composants sont coordonnés et les deux dépendent de la tête du syntagme. Les composants de telles

formations sont autonomes, voire quasiment toujours liés par un trait d'union qui paraît servir ici de marque syntaxique plutôt que morphologique. Nous avons attribué à ces formations l'étiquette de **quasi-composés**. Les composants sont attestés dans les dictionnaires et les corpus, et ils ne sont pas multilingues.

La classification proposée essaie de capturer la variété de la nature et du degré d'autonomie des composants à travers des ressources lexicales (dictionnaires, corpus, lexiques complémentaires comme celui de racines néoclassiques ou de préfixes) pour identifier les composants à l'intérieur d'une unité graphique.

5.1.6 Analyse des composés morphologiques

Après avoir indiqué les phénomènes que nous allons traiter, nous examinons maintenant quelle analyse nous allons mettre en œuvre pour ces phénomènes. Nous introduisons quelques définitions complémentaires, relatives à la segmentation des composés.

Base, radical, composant

Nous adoptons les définitions suivantes de **base**, **radical**, **composant** et **lemme du composant**.

Base et radical. D'après Gardes-Tamine (2010, p. 58), « lorsqu'on enlève un affixe à un mot, on obtient la **base** sur laquelle il est formé : la base à laquelle s'adjoint *ir-* dans *irréalisable* est *réalisable*. La base à laquelle s'adjoint *-able* dans ce dernier mot est *réalis-*. Lorsque tous les affixes ont été ôtés, il reste une base minimale, qui coïncide avec un morphème, que l'on appelle **radical**. »

Composant et son lemme. Nous appelons **composant** une sous-partie d'un composé morphologique qui correspond à un élément sémantique, mais pas nécessairement à un morphème libre. Un composé est divisible par un nombre entier de composants (à l'exception du tiret que nous n'allons pas compter comme une partie du composant). Ainsi, le composé EN *toolbar* a des composants : *tool* et *bar* ; le composé DE *Staatsfeind* 'ennemi public' a des composants *Staats* et *Feind*.

Puisqu'un composant ne correspond pas toujours à une unité lexicale autonome, nous devrons restaurer sa forme de base, celle qui correspondrait à une entrée du dictionnaire, le **lemme du composant**. Ainsi, dans l'exemple DE *Staatsfeind* le lemme du composant *Staats* est le nom *Staat* 'état', et le lemme du composant *Feind* est le nom *Feind* 'ennemi'. Pour les éléments néoclassiques, le lemme est postulé comme toujours égal au composant.

Notation. Pour l'analyse des composés, nous utilisons la représentation suivante :
<composé> = <lemme du composant 1> + <lemme du composant 2> (+ ... + <lemme du composant N>), par exemple :

FR mammographie = mammo + graphie.

Pour les composés d'autres langues que le français, nous rajoutons la traduction entre cotes simples :

DE Staatsfeind 'ennemi public' = Staat 'état' + Feind 'ennemi'.

Pour les composés russes, nous citons le mot en cyrillique une fois, mais l'analyse sera donnée en translittération pour faciliter la compréhension :

RU ветрогенератор 'générateur éolien'
vetrogenerator = veter 'vent' + generator 'générateur'.

Parfois nous avons besoin de montrer clairement les composants, et pas seulement leurs lemmes. Dans de tels cas nous utilisons le signe des deux points comme séparateur entre le composant et son lemme : <composant 1>:<lemme du composant 1> '<traduction>'

vetrogenerator = vetro:veter 'vent' + generator:generator 'générateur'.

Allomorphes et éléments de liaison

Les mécanismes de composition sont plus ou moins complexes en fonction des langues. Dans les langues plutôt analytiques comme les langues française et anglaise les composants sont généralement concaténés : FR *kilowatt-heure*, EN *parrotfish* 'poisson perroquet'. Voilà pourquoi dans la tradition anglophone on parle parfois de « *root compounding* », composition des radicaux.

Dans les langues synthétiques, les composants ne correspondent pas toujours exactement à leurs lemmes :

DE Staatsfeind 'ennemi public' = Staat 'état' + Feind 'ennemi';

RU водохранилище 'réservoir d'eau' :
vodohraniliše = voda 'eau' + hraniliše 'réservoir'.

Un ou plusieurs caractères divisent souvent les bases. Ceux-ci peuvent être des morphèmes casuels, comme dans l'exemple

RU пятиметровый 'de cinq mètres' :
pâtimetrovyj = pâti:pât'$_{NUM.GEN}$ 'cinq' + metr:metr 'mètre' + ov$_{SUF}$ + yj$_{DES}$.

D'autres composants ne correspondent à aucune des formes dans le paradigme de leur lemme, cf. RU *vodo* dans l'exemple (*vodohraniliše*), idem. pour la langue grecque (Ralli, 2013).

Pour l'allemand, la question se pose si les composants sont des formes paradigmatiques. D'un côté, un composant DE peut correspondre à une des formes fléchies dans le paradigme de son lemme (forme plurielle ou cas génitif), cf. dans l'exemple DE *Staatsfeind*, *Staats* coïncide avec le génitif de *Staat*. De l'autre côté, la sémantique des composants ne correspond pas toujours à la sémantique du pluriel ou du génitif (Langer, 1998).

Ces éléments tels que "o" et "e" en russe, "(e)s", "(e)n" en allemand, etc. sont appelés différemment en fonction de la langue et de l'approche, i.e. « voyelle de liaison » en russe (RU « соединительная гласная »), « élément de liaison » en allemand (« *Fugenelement* »), « morphème de liaison », « interfixe » (Mel'čuk, 1997) , etc.

Il n'y a pas non plus d'accord général sur la nature de ces éléments. Les deux interprétations les plus répandues sont les suivantes :

1. Ce sont des interfixes, i.e. des affixes interradicaux propres aux langues fusionnelles (Dressler et Barbaresi, 1994; Mel'čuk, 1997);
2. Ce ne sont pas des morphèmes parce qu'ils sont sémantiquement vides, ces éléments font partie d'un allomorphe du premier radical (Plungian, 2000; Booij, 2005). Parfois le terme « base compositionnelle » (« *compounding stem* ») (Langer, 1998) est utilisé pour distinguer cet allomorphe des autres.

Dans ce travail, nous allons nous en tenir à la deuxième interprétation car ce cadre nous paraît plus adapté pour décrire comment il est possible de restaurer son lemme à partir du composant. Cela n'empêche pas de reconnaître l'existence d'éléments de liaison dans les langues synthétiques, même si leur statut morphématique est contestable.

Outre l'insertion d'un élément de liaison, l'alternance dans la base est possible dans certains allomorphes :

DE Gänseklein 'abats d'oie' = Gans 'oie' + klein 'petit';
RU ветрогенератор 'générateur éolien'
vetrogenerator = veter 'vent' + generator 'générateur'.

Nombre de composants

Un composé peut contenir plus de deux composants, par exemple : EN *histopathology* = *histo* + *patho* + *logy* ou DE *Restlebenszeitrisiko* = *Rest* + *Leben* + *Zeit* + *Risiko* 'temps moyen de survie'. Il est cependant possible de traiter la composition, en tant que processus, comme la combinaison de deux parties. Benveniste (1974, p. 146) a défendu ce postulat :

> « Nous posons en principe qu'un composé comporte toujours et seulement deux termes. <...> Mais, des deux termes d'un composé, l'un peut être lui-même composé. <...> Le composé devenant terme de composé compte pour un seul terme; il n'y en a toujours que deux dans le composé nouveau. »

Selon cette position, le mot *histopathology* doit être analysé en deux étapes :

> histopathology = histo + pathology
> pathology = patho + logy

Nous partageons ce principe, et nous allons appliquer par la suite cette analyse récursive.

5.1.7 Hypothèses par rapport à la segmentation

En guise de résumé, nous adoptons les hypothèses suivantes que nous mettrons en œuvre dans le chapitre 6:

1. Au moment de la composition, deux et seulement deux composants sont impliqués.
2. La composition est récursive, un composant peut être lui-même formé par la composition.
3. Un composé en tant que résultat de la composition peut donc comporter deux ou plusieurs composants.
4. La frontière entre les composants peut être plus ou moins explicite : marquée par le tiret, marquée par un élément de liaison ou non marquée.
5. Les composants ne correspondent pas toujours exactement à leurs lemmes. Des règles peuvent être formulées pour décrire comment le lemme peut être restauré à partir d'un composant dans une langue donnée. Certains allomorphes sont cependant des variantes irrégulières et ne pourront pas être restaurés à l'aide des règles.

5.2 Segmentation automatique des composés.

Après avoir délimité les phénomènes traités dans ce travail et après avoir formulé les hypothèses par rapport à la segmentation des composés, nous résumons dans cette partie les difficultés rencontrées lors de la segmentation de mots composés. Nous examinons des solutions envisagées, décrites dans la littérature et implémentées dans des outils, ainsi que les moyens d'évaluer la qualité de segmentation.

5.2.1 Difficultés liées au traitement des composés

Le traitement des mots composés comporte plusieurs difficultés pour les systèmes du TAL. Tout d'abord, comment savoir si un mot est composé ou non? Le trait d'union peut servir de repère, mais nombreux sont les composés « soudés ». Le simple filtrage avec un dictionnaire afin d'éliminer les non-composés n'est pas suffisant parce que tous les mots n'y apparaissent pas, notamment les noms propres et les dérivés de mots présents dans le dictionnaire (adjectifs dénominaux, diminutifs, etc.).

Si le mot est composé, pour le segmenter il faut trouver l'endroit où se situe la frontière (ou les frontières) entre les composants. Le découpage est parfois ambigu lorsque plusieurs composants candidats existent en tant que mots pleins. Ainsi, le mot allemand *Traktionsbatterie* 'batterie de traction' peut être segmenté de deux manières :

Traktionsbatterie →
traktion + batterie
trakt + ion + batterie.

Ici, seule la première analyse est correcte.

Le niveau de difficulté le plus élevé pour les systèmes automatiques est atteint pour des langues dans lesquelles plusieurs segmentations homonymiques sont possibles pour un mot composé. Ainsi le mot suédois *bildrulle* a également deux segmentations (l'exemple tiré de (Sjöbergh et Kann, 2004)) :

bildrulle →
bild + rulle 'bobine de film'
bil + drulle 'mauvais conducteur'.

La bonne interprétation dépend du contexte.

Enfin, même si les frontières sont bien définies, il faut aussi trouver les lemmes de tous les composants, étant donné que les composants ne sont pas toujours autonomes. Certains systèmes du TAL optent pour le stockage dans le lexique de tous les composants connus avec leurs lemmes. Cette solution nous semble cependant insatisfaisante pour des tâches multilingues car la construction manuelle d'un tel lexique est un travail extrêmement complexe et ne résout les problèmes que partiellement car de nouveaux composés apparaissent constamment dans les textes.

Nous répertorions ci-dessous quelques approches existantes de la segmentation automatique des composés. On peut distinguer les méthodes qui s'appuient entièrement sur des connaissances linguistiques spécifiques à une langue, des méthodes basées sur l'utilisation de corpus, ou encore des méthodes probabilistes.

5.2.2 Méthodes basées sur des connaissances linguistiques

Les méthodes de ce premier type utilisent des connaissances spécifiques à une langue traitée qui peuvent être définies sous forme de grammaire ou de règles.

L'analyseur morpho-sémantique pour la langue française DériF (Namer, 2003) traite les mots morphologiquement construits, entre autres les composés natifs et néoclassiques. L'analyse s'appuie sur un lexique des bases, une table de correspondances entre les bases autonomes et les éléments néoclassiques, une liste d'affixes (préfixes, suffixes) et des règles de construction morphologique avec certaines restrictions sur la compatibilité entre les éléments. L'analyseur s'applique à des formes lemmatisées et accompagnées de leur catégorie grammaticale.

Dans le même paradigme, l'analyseur morphologique pour l'allemand SMOR (Schmid et al., 2004) est capable de traiter les composés grâce à la modélisation de grammaire et le lexique des bases (y compris des bases compositionnelles). Des restrictions s'appliquent afin de filtrer les formations irréalistes, de manière à ce que seules des analyses vraisemblables soient générées. Un point faible est qu'un composé ne sera pas segmenté si au moins une de ses bases manque dans le lexique. De plus, dans le cas où plusieurs segmentations sont possibles, elles ne sont pas ordonnées. SMOR traite à la fois la composition et la morphologie (dérivationnelle et flexionnelle), ce qui permet de décomposer des formes lemmatisées ainsi que des formes fléchies.

Le segmenteur BananaSplit (Ott, 2005) effectue la segmentation des composés allemands en s'appuyant sur un dictionnaire. Il utilise aussi des règles de transformation des composants définies à partir de l'étude de Langer (1998). Des exemples de règles sont :

— "s" → "": Staatsfeind 'ennemi public' = Staat 'état' + Feind 'ennemi';
— "n" → "": Soziologenkongreß = Soziologe 'sociologue' + Kongreß 'congrès';
— "" → "e": Kirchturm = Kirche 'église' + Turm 'tour'.

BananaSplit parcourt un mot de la droite vers la gauche et compare toutes les sous-chaînes possibles avec les entrées du dictionnaire. Quand il a trouvé la sous-chaîne la plus longue existant dans le dictionnaire, il vérifie si la partie restante y appartient aussi (en appliquant des règles là où c'est possible). Si cette recherche aboutit, les mots trouvés dans le dictionnaire sont considérés comme les lemmes des composants. Parfois plusieurs règles peuvent être appliquées à un composant ce qui génère plusieurs candidats lemmes. Dans ce cas le programme calcule les similarités des chaînes de caractères (la distance de Levenshtein) entre le composant avant l'application des règles et les candidats lemmes : le lemme avec la similarité maximale est retenu. Cette méthode ne nécessite pas de grammaire spécifique à une langue, ni de liste des bases formant des composés, lesquelles sont difficiles à construire pour une nouvelle langue. Par contre, elle peut produire des segmentations irréalistes :

**Blindleistungseinspeisung* 'alimentation en puissance réactive'
= *Blindleistung* 'puissance réactive' + *sein* 'être' + *speisung* 'alimentation'

La segmentation correcte est *Blindleistung* 'puissance réactive' + *Einspeisung* 'alimentation'. Un système basé sur la grammaire ne produit pas de telles erreurs grâce aux restrictions (le verbe *sein* n'est pas une base compositionnelle). L'auteur évalue son segmenteur sur un ensemble de mots non-lemmatisés (Ott, 2006). Pour traiter la flexion, il introduit une liste de désinences. BananaSplit ne segmente que les mots qui n'appartiennent pas au dictionnaire employé. Tous les mots répondant à ce critère et dont les composants sont trouvés dans le dictionnaire sont considérés comme composés.

5.2.3 Méthodes basées sur l'utilisation des corpus

Le travail pionnier utilisant les statistiques de corpus pour la segmentation a été effectué par Koehn et Knight (2003) sur des composés allemands. Pour un mot donné, l'algorithme génère toutes les segmentations possibles (également en tenant compte des éléments de liaison) et estime la probabilité de chaque segmentation (S) à partir de la moyenne géométrique des fréquences absolues des composants (p_i) dans le corpus monolingue :

$$score(S) = (\prod_{i=1}^{n} freq(p_i))^{\frac{1}{n}} \tag{5.1}$$

où n renvoie au nombre de composants. Seuls les noms, les adverbes, les adjectifs et les verbes sont considérés comme des composants. La segmentation candidate obtenant le meilleur score est choisie comme la bonne analyse. Le mot initial non-segmenté est considéré lui aussi comme un candidat, avec un score égal à sa fréquence dans le corpus. Cette approche permet de segmenter un composé même si ses composants sont trop récents ou trop spécialisés et ne sont pas recensés dans un dictionnaire. Elle ne dépend pas d'une grammaire spécifique à une langue et peut donc être appliquée à différentes langues. Des connaissances linguistiques telles que l'insertion des éléments de liaison peuvent toutefois y être intégrées. Quant aux points faibles de la méthode, des segmentations irréalistes sont parfois générées. Une erreur fréquente est la sous-segmentation (« *under-splitting* »), car la fréquence d'un mot composé courant peut s'avérer plus élevée que la moyenne géométrique de ses parties.

Les auteurs utilisent la fréquence absolue d'un composant « tel quel » et non de son lemme, c'est-à-dire que les formes DE *Plan* et $Planes_{N.GEN}$ ont des fréquences différentes, égales au nombre d'apparitions de chaque forme dans le corpus. Cela disperse la fréquence du lexème selon les formes non-homonymiques de son paradigme. En revanche, cette approche permet d'analyser les composés apparus dans le corpus en forme fléchie (par exemple, $Aktionsplanes_{N.GEN}$ 'plan d'action').

En vue d'une application de traduction et non exclusivement de segmentation, Koehn et Knight (2003) ont ajouté une étape supplémentaire nécessitant un corpus parallèle : les composants candidats sont d'abord

traduits en anglais séparément (en supposant que les composés allemands sont généralement traduits en anglais par des composés syntagmatiques), et ensuite pour chaque segmentation le nombre de correspondances entre les traductions des composants et des « vraies » traductions qui apparaissent dans la partie anglaise du corpus est calculé. La bonne segmentation sera le candidat avec le nombre de correspondances le plus important. Le composé est ensuite remplacé dans les textes sources par ses composants. Cette méthode s'avère particulièrement robuste, mais son utilisation est possible sous réserve de posséder des corpus parallèles et alignés au niveau des mots ou un ensemble de phrases parallèles, ressources qui ne sont pas disponibles pour toutes les paires de langues et tous les domaines de spécialité.

La méthode (Koehn et Knight, 2003) dans sa variante monolingue a été utilisée dans de nombreux travaux sur la segmentation en allemand ou dans d'autres langues, avec des améliorations variées. Stymne (2008) et Stymne et al. (2013) appliquent la même méthode en utilisant la moyenne arithmétique au lieu de la moyenne géométrique pour contourner le problème de sous-segmentation. En outre, les auteurs rajoutent la restriction suivante : la tête du composé (le composant droit) doit appartenir à la même catégorie grammaticale que le composé entier, ce qui améliore la précision de la segmentation. Les expériences rapportées dans (Stymne et al., 2013) concernent trois langues germaniques : l'allemand, le suédois et le danois. La considération que la tête du composé correspond à son composant droit qui est vraie pour les langues germaniques n'est cependant pas universelle et ne peut pas être systématisée à toutes les langues.

Weller et Heid (2012) emploient la méthode (Koehn et Knight, 2003) couplée avec la restriction sur la catégorie grammaticale du composant droit. En sortie, leur segmenteur[4] fournit les meilleures segmentations avec une catégorie grammaticale pour chaque composant:

DE Dampfturbine$_N$ - Dampf$_N$ Turbine$_N$ 'turbine à vapeur'

Cette information peut servir pour la traduction vers d'autres langues où les composés morphologiques sont traduits plutôt par des composés syntagmatiques. Les auteurs manipulent les fréquences des lexèmes et non des composants, c'est-à-dire qu'ils regroupent les fréquences des formes paradigmatiques (DE *Plan, Planes, Pläne*). Afin de pouvoir segmenter un composé en forme fléchie, ils utilisent une liste des toutes les formes apparues dans le corpus accompagnées de leurs lemmes. La méthode a été appliquée à la segmentation des termes composés allemands relatifs à un domaine de spécialité.

Fritzinger et Fraser (2010) proposent de combiner l'analyse linguistique avec celle basée sur le corpus. Pour l'allemand, ils utilisent les segmentations linguistiquement motivées produites (mais pas ordonnées) par l'analyseur morphologique SMOR (Schmid et al., 2004), et ils les ordonnent d'après les scores attribués par la méthode de Koehn et Knight (2003). Cette combinaison permet d'éviter les segmentations irréalistes (*Blindleistung + sein + Speisung*) et d'ordonner les candidats. La qualité de segmentation repose sur l'analyseur morphologique.

5.2.4 Méthodes probabilistes

Les méthodes probabilistes de reconnaissance et de segmentation des mots composés sont entièrement indépendantes de la langue. Contrairement aux méthodes basées sur l'utilisation du corpus, elles s'appuient sur la probabilité bayésienne des composants, et non sur la probabilité empirique égale au nombre d'apparitions des composants dans le corpus.

Creutz et Lagus (2002) ont proposé une méthode automatique d'acquisition d'un lexique des *morphes* d'une langue à partir de textes non-annotés (il s'agit donc d'une méthode non-supervisée). Les morphes dans ce cadre théorique sont des réalisations en surface des morphèmes, unités minimales et indivisibles porteuses du sens. Ce lexique comprend des unités lexicales employées indépendamment dans les textes, mais aussi des segments non-autonomes qui apparaissent à plusieurs reprises en tant que sous-parties de mots différents. La méthode permet donc d'effectuer l'analyse morphologique des mots simples, ainsi que

4. http://www.ims.uni-stuttgart.de/~wellermn/tools.html

la segmentation d'un composé en composants. La méthode a été implémentée dans l'analyseur morphologique Morfessor. La première version de Morfessor s'appuyait sur le modèle statistique intitulé « *Minimum Description Length* » (*longueur minimale de description*), qui a été remplacé dans la deuxième version par un modèle équivalent *maximum a posteriori*. L'outil a été conçu pour les langues ayant une morphologie concaténative, et il produit ainsi la segmentation sans lemmatisation des composants, e.g. DE *Kirchturm* sera segmenté en *Kirch* + *Turm* et pas en *Kirche* + *Turm*. Il existe une variante de cette méthode, Allomorfessor (Virpioja et al., 2009), conçue pour prendre en compte le phénomène d'allomorphie grâce au calcul de similarité de chaînes de caractères. Les deux variantes sont capables de traiter des composés en forme fléchie car l'objectif de Morfessor est la segmentation en morphes, qui peut comprendre des affixes flexionnels. L'évaluation a été effectuée sur les langues anglaise et finnoise (et aussi sur l'allemand, le turc et l'arabe dans le cadre de la campagne Morpho Challenge, cf. section 5.2.5).

Dyer (2009) applique le modèle de l'*entropie maximale* pour la segmentation des composés. Sa méthode génère les treillis de segmentation qui contiennent plusieurs segmentations candidates pour un composé. Afin de construire un treillis de segmentation pour un nouveau mot, l'auteur exploite plusieurs caractéristiques (la longueur des composants, la fréquence des composants dans un corpus en tant que mot indépendant, les premiers caractères, etc.) ainsi que leur poids appris pendant l'entraînement préalable du modèle sur un certain nombre de treillis de référence construits manuellement (c'est donc une méthode supervisée). L'algorithme a été testé pour les langues allemande, hongroise et turque, tandis que les treillis de référence ont été utilisés seulement pour l'allemand : pour les deux autres langues, les mêmes poids ont été appliqués. L'auteur démontre que les poids appris pour une langue peuvent être utilisés pour d'autres langues. Ce modèle permet l'intégration de certaines connaissances linguistiques telles que les éléments de liaison et un anti-dictionnaire des mots ne formant jamais de composés. L'intégration des éléments de liaison permet d'effectuer dans une certaine mesure la lemmatisation des composants.

Hewlett et Cohen (2011) détectent automatiquement la place des frontières de composants. La méthode a pour objectif final l'établissement des limites non seulement entre les parties des mots composés, mais entre tous les mots d'un corpus. Le corpus est donc représenté sous la forme d'une longue chaîne de caractères sans aucune segmentation préalable ou seulement avec la segmentation en phrases. La méthode emploie le modèle de *longueur minimale de description* couplée avec des algorithmes basés sur l'entropie. L'entropie est utilisée comme une mesure de la probabilité d'apparition d'une séquence de caractères dans une langue : l'entropie à l'intérieur d'un mot est relativement basse, tandis que l'entropie sur les frontières des mots est assez élevée. Cet algorithme peut servir pour la « *tokenisation* » des textes (c'est-à-dire l'établissement des frontières entre les mots, pris dans un sens plus large que le mot graphique), y compris la tokenisation des textes en langues polysynthétiques, et entre autres pour la segmentation des mots composés en langues où la composition est fréquente et plutôt concaténative car la méthode ne gère pas l'allomorphie. Les auteurs ont appliqué leur méthode à la segmentation des composés en anglais, chinois et thaï.

Macherey et al. (2011) proposent un algorithme non-supervisé pour extraire automatiquement des opérations morphologiques sur les frontières de composants. Des exemples d'opérations extraites pour l'allemand sont : "-"/ε , "s"/ε , "es"/ε , "n"/ε , "e"/ε , "en"/ε (ε signifiant une chaîne vide). Afin d'entraîner un modèle pour une nouvelle langue, ils exploitent les tables de traduction vers l'anglais. Cela permet d'extraire des composés et leurs composants pour en déduire les opérations fréquentes dans cette langue. Pour distinguer les mots composés d'autres mots inconnus du dictionnaire, Macherey et al. (2011) compilent la liste des non-composés de manière automatique en utilisant également les tables de traduction. La méthode a été testée pour plusieurs langues : danois, allemand, norvégien, suédois, grec, estonien, finnois.

La variété des langues et des données utilisées pour les expériences rendent difficile la comparaison des performances de plusieurs méthodes de segmentation en s'appuyant seulement sur les résultats rapportés dans les publications. De plus, il existe des manières différentes d'évaluer la qualité de la segmentation, ce que nous examinons dans la section suivante.

5.2.5 Évaluation de la qualité de segmentation

L'évaluation de la performance d'une méthode de segmentation automatique peut être effectuée comme une tâche indépendante (*évaluation intrinsèque*) ou via une autre application du TAL (*évaluation extrinsèque*).

Dans le scénario de **l'évaluation intrinsèque**, la sortie du système peut être soit annotée manuellement par un expert, soit comparée à un ensemble de données de référence déjà annotées, exactement comme nous l'avons vu pour l'évaluation de l'extraction terminologique. Quelle que soit la procédure d'annotation, pour mesurer le résultat quantitativement, les mesures traditionnelles telles que la précision, le rappel et la F-mesure peuvent être appliquées ensuite. La précision montre généralement combien de segmentations sont correctes parmi celles réalisées par le système, tandis que le rappel montre combien parmi les composés analysés sont segmentés correctement par le système. Quant à la F-mesure, elle combine la précision et le rappel. Cependant nous pouvons observer des différences dans l'usage de ces mesures pour la segmentation.

Pour expliquer ces différences sur l'exemple de l'évaluation utilisant des données de référence, nous introduisons les variables suivantes :

U_C - unités candidates, i.e. unités produites par le système évalué;
U_R - unités de référence, i.e. unités annotées qui forment l'ensemble de référence.

Ainsi, la précision P est calculée comme la taille de l'intersection entre les unités candidates et les unités de référence, divisée par le nombre d'unités candidates :

$$P = \frac{|U_C \cap U_R|}{|U_C|} \tag{5.2}$$

Le rappel R est calculé comme la taille de l'intersection entre les unités candidates et les unités de référence, divisée par le nombre d'unités de référence :

$$R = \frac{|U_C \cap U_R|}{|U_R|} \tag{5.3}$$

La F-mesure balancée est calculée comme la moyenne harmonique entre les deux mesures :

$$F = \frac{2 \times P \times R}{P + R} \tag{5.4}$$

Dans les travaux de Ott (2006); Koehn et Knight (2003) et d'autres auteurs utilisant les méthodes fondées sur le corpus, les unités candidates sont des segmentations produites par le système, et les unités de références sont des mots graphiques annotés avec leurs segmentations correctes ou avec l'étiquette de non-composés.

Les mêmes mesures peuvent également être calculées sur les frontières entre les composants (Hewlett et Cohen, 2011; Virpioja et al., 2013) (dans ce cas, les unités candidates sont des frontières mises par le système, et les unités de référence sont des frontières marquées dans les données de référence) ou sur les composants identifiées (Hewlett et Cohen, 2011). Pour donner un ordre de grandeur, Hewlett et Cohen (2011) obtiennent une F-mesure sur les frontières qui varie entre 83 % et 93 %, et une F-mesure sur les composants entre 58 % et 80 % (en fonction de la langue et du corpus). Virpioja et al. (2013) avec Morfessor obtiennent sur les frontières des composants une F-mesure d'environ 76 % pour l'anglais et d'environ 57 % pour le finnois.

Au lieu de la F-mesure, Koehn et Knight (2003) introduisent une autre mesure propre à la tâche de segmentation, « *accuracy* », que nous traduisons comme *exactitude*. L'exactitude montre sur l'ensemble des mots à analyser (composés et non-composés), combien d'analyses sont correctes (c'est-à-dire les cas où les mots qui doivent être segmentés sont segmentés correctement par le système, et ceux où les mots qui ne doivent pas être segmentés ne le sont effectivement pas). Pour redéfinir l'exactitude en termes de U_C et U_R, nous avons besoin d'une autre variable U_N :

U_N - unités non-segmentées, i.e. unités qui n'ont pas été segmentées par le système évalué.

L'exactitude A est alors calculée ainsi :

$$A = \frac{|U_C \cap U_R| + |U_N \cap U_R|}{|U_C| + |U_N|} \tag{5.5}$$

Nous remarquons que $|U_C| + |U_N|$ est égale à $|U_R|$. Sans utilisation du corpus parallèle anglais, Koehn et Knight (2003) obtiennent pour l'allemand une exactitude d'environ 96 %, et avec le corpus parallèle, l'exactitude remarquable de 99 %. Cette mesure a été réutilisée dans d'autres travaux, également pour l'allemand. Sur les mots non-lemmatisés en entrée, Ott (2006) rapporte avec BananaSplit une exactitude de 74 %. Avec la moyenne arithmétique des fréquences, Stymne et al. (2013) atteignent l'exactitude de 92 %. Un résultat très proche est obtenu par Fritzinger et Fraser (2010) avec une méthode mixte combinant la sortie de SMOR et la moyenne géométrique des fréquences.

L'évaluation extrinsèque montre à quel point l'utilisation d'une méthode influence les résultats d'une autre application plus large. Elle s'effectue en utilisant la sortie d'un système comme une des entrées de l'application remplissant une tâche pour laquelle il existe des méthodes standardisées d'évaluation. La segmentation a été évaluée via la traduction automatique statistique (« *statistical machine translation* ») en comparant la qualité de traduction faite sans segmentation et avec segmentation des composés (Koehn et Knight, 2003; Dyer, 2009; Macherey et al., 2011; Stymne et al., 2013) à l'aide de BLEU (cf. annexe A). Pour la traduction d'une langue compositionnelle vers l'anglais, tous les auteurs ont constaté une amélioration de la qualité de traduction après segmentation qui varie entre 0,6 et 4 points en termes de BLEU.

Les résultats d'évaluations intrinsèque et extrinsèque ne coïncident pas toujours. Ainsi, Stymne et al. (2013) démontre sur l'exemple de la langue allemande que dans l'évaluation de segmentation en tant que tâche indépendante, la moyenne géométrique des fréquences des composants donne de meilleurs résultats en termes de précision et de rappel, tandis qu'en étant évaluée dans le cadre de la traduction automatique, la segmentation utilisant la moyenne arithmétique permet d'obtenir un meilleur score BLEU.

La différence dans la manière d'évaluer les résultats complique la comparaison des méthodes. Le meilleur moyen reste de réévaluer les systèmes sur la base d'un même jeu de données. Toutefois, peu de systèmes sont disponibles au téléchargement, de même que les données utilisées pour leur évaluation.

Il existe des campagnes d'évaluation des méthodes non-supervisées et semi-supervisées d'apprentissage de la morphologie, telle que Morpho Challenge [5] qui fournit les données standardisées en plusieurs langues (anglais, allemand, finnois, turc, arabe). Voici quelques exemples d'analyses pour EN (nous gardons ici la notation utilisée dans la campagne) :

abusing ab:ab_p us:use_V ing:+PCP1
adversaries advers:adverse_A ari:ary_s es:+PL
airline air:air_N line:line_N

Cette campagne vise l'analyse morphologique complète, qui comprend les morphèmes flexionnels et dérivationnels, et pas seulement les composants des mots composés.

5. http://research.ics.aalto.fi/events/morphochallenge/

5.3 Bilan

Cet aperçu des méthodes existantes nous permet de constater que la segmentation des composés est une tâche importante pour le traitement de nombreuses langues, et dans le même temps qu'il s'agit d'une tâche complexe pour laquelle les solutions pratiques proposées fonctionnent plus ou moins bien selon la langue.

Parmi les difficultés importantes pour les systèmes automatiques se trouve le phénomène d'allomorphie des composants. Pour traiter l'allomorphie, une solution possible est d'utiliser des connaissances linguistiques spécifiques à une langue. Cela semble aller à l'encontre de la tendance à demeurer indépendant de la langue. Pourtant ces deux approches ne sont pas réellement opposées. Les méthodes indépendantes de la langue intègrent de plus en plus des éléments de connaissances linguistiques, et les résultats des deux méthodes peuvent aussi être combinés pour produire un résultat final de meilleure qualité.

Nous avons voulu trouver une méthode de reconnaissance et de segmentation des composés qui serait multilingue, adaptable à la langue et au domaine de spécialité mais aussi capable de traiter l'allomorphie des composants. Parmi les implémentations disponibles des méthodes énumérées, aucune ne répond à tous ces critères. Deux types de méthodes s'en rapprochent : les méthodes basées sur la fréquence des composants dans le corpus, et les méthodes d'acquisition automatique du lexique des composants à partir de corpus. Les premières nécessitent un inventaire des éléments de liaison pour chaque langue sans lequel elles ne pourront pas traiter les allomorphes. Les dernières ont pour but l'analyse morphologique complète et non l'analyse des composés; elles ne permettent pas d'intégrer des connaissances linguistiques sur les éléments de liaison. Aucune des méthodes, en l'état, ne prévoit l'adaptation au domaine de spécialité. Nous allons proposer notre méthode en essayant d'adopter les points forts des méthodes examinées et de remédier, au moins en partie, à leurs inconvénients.

6

Segmentation multilingue des composés

Après avoir examiné les méthodes principales existant pour le traitement des mots composés et identifié leurs points forts et faibles, nous allons proposer notre méthode conçue, en premier lieu, pour le traitement des termes composés issus des domaines spécialisés. Cette méthode combine l'utilisation de corpus et de dictionnaires et, optionnellement, de ressources lexicales complémentaires. Elle peut être adaptée à la langue et au domaine de spécialité. La méthode est supervisée, elle nécessite l'apprentissage des paramètres pour une nouvelle langue.

Dans ce chapitre, nous présentons cette méthode et son implémentation (section 6.1), décrivons les expériences menées et évaluons les résultats obtenus, d'un point de vue quantitatif et qualitatif (section 6.2). Nous finissons par comparer notre méthode avec d'autres approches de l'état de l'art (section 6.3).

6.1 CompoST : méthode d'identification et de segmentation des termes composés

6.1.1 Introduction

Le traitement des composés nécessite une méthode générique, c'est-à-dire qui pourrait être appliquée à différentes langues grâce aux caractéristiques indépendantes de la langue, et suffisamment robuste pour fonctionner dans le cas d'une nouvelle langue sans nécessiter de connaissances préalables. Néanmoins si pour une langue donnée nous disposons de connaissances spécifiques, cette méthode doit être capable de les intégrer pour optimiser les résultats.

De l'état de l'art examiné dans la section 5.2 nous retenons quelques points importants. Premièrement, les composants ou leurs lemmes sont attestés soit dans un dictionnaire, soit dans un corpus de textes. L'utilisation des corpus se montre bénéfique pour pallier à l'incomplétude des dictionnaires. L'utilisation de corpus assure aussi l'indépendance de la langue traitée, le corpus étant une ressource facilement constructible pour une langue. Les deux types de méthodes les plus proches de notre but s'appuient sur un corpus. Dans le même temps, le dictionnaire construit par des experts de la langue reste une ressource plus « sûre » et moins bruitée qu'un corpus compilé généralement de manière automatique ou semi-automatique. C'est également une ressource disponible pour de nombreuses langues. Notre méthode va donc combiner l'exploitation de ces deux types de ressources : corpus et dictionnaires.

Dans notre optique de traitement des termes spécialisés, le recours à un guidage provenant du corpus s'impose également pour prendre en compte le domaine de spécialité. Pour un tel guidage nous choisissons la spécificité des lexèmes à l'intérieur d'un domaine, calculée à partir d'un corpus spécialisé et d'un corpus général.

Le deuxième point important est la capacité à traiter les allomorphes des racines formant un composé. Certaines méthodes utilisent pour cela des règles spécifiques pour la langue traitée, ce qui permet d'obtenir une bonne précision pour cette langue mais nécessite de telles règles pour chaque nouvelle langue. Une autre solution (notamment réalisée dans Allomorfessor (Virpioja et al., 2009)) qui permet de conserver l'indépendance par rapport à la langue est d'utiliser le calcul de similarité des chaînes de caractère entre les composants et des unités lexicales autonomes. Dans notre méthode nous utilisons également ce calcul pour assurer le caractère générique de la méthode, et nous intégrons l'usage des règles linguistiques sous forme d'option pour que la méthode soit adaptable à la langue. Même pour les langues pour lesquelles nous pouvons formuler des règles, le calcul de similarité de chaînes pourrait être utile si les règles ne sont pas exhaustives.

Nous proposons une stratégie unifiée pour traiter tous les phénomènes énumérés dans notre classification opérationnelle (section 5.1.5), à savoir les composés natifs, néoclassiques, empruntés, ainsi que les mots préfixés et les quasi-composés. La liste des éléments néoclassiques et des préfixes pourra être utilisée comme une ressource complémentaire au dictionnaire pour segmenter des néoclassiques et des préfixés.

Enfin, pour équilibrer l'influence de chaque constituant du système (l'utilisation de la spécificité, des ressources lexicales, des règles, de la similarité) nous introduisons la phase d'apprentissage nécessaire pour optimiser les paramètres de segmentation pour la langue traitée et pour les ressources exploitées.

Pendant la phase de segmentation, un mot composé sera segmenté à chaque étape en deux parties en accord avec nos hypothèses formulées dans la section 5.1.7. Le processus étant récursif, le nombre de composants dans l'analyse finale est donc théoriquement borné par le nombre de caractères (et en pratique, nous allons le fixer pour chaque langue séparément).

Le tour d'horizon des études réalisées sur plusieurs langues et dans des contextes applicatifs différents nous a amené à réfléchir également au format d'entrée adapté pour notre système. Certaines méthodes sont appliquées aux formes fléchies, et d'autres, aux mots lemmatisés. Le fait de traiter les composés lemmatisés permet de réduire considérablement la taille du lexique du corpus, particulièrement pour des langues ayant un système de flexion substantivale très riche. Nous choisissons donc cette option.

L'algorithme proposé est présenté schématiquement sur la figure 6.1. Son implémentation en JAVA, nommée CompoST (« *Compound Splitting Tool* »), est accessible en ligne, ainsi que des jeux de données [1].

6.1.2 Segmentation du composé

Pour segmenter un composé, nous commençons par générer toutes ses segmentations possibles en deux parties, de taille supérieure ou égale à la longueur minimale acceptée pour un composant : nous avons adopté une longueur minimale de 3 caractères, ce qui correspond à la longueur constatée dans l'état de l'art pour la segmentation automatique, en dessous de laquelle la décomposition produit plus d'analyses erronées que correctes (Koehn et Knight, 2003; Dyer, 2009).

DE *Traktionsbatterie* ('batterie de traction') :
traktionsbatterie → tr + aktionsbatterie
traktionsbatterie → tra + ktionsbatterie
...
traktionsbatterie → traktion + batterie
...
traktionsbatterie → traktionsbatter + ie

1. https://logiciels.lina.univ-nantes.fr/redmine/projects/compost

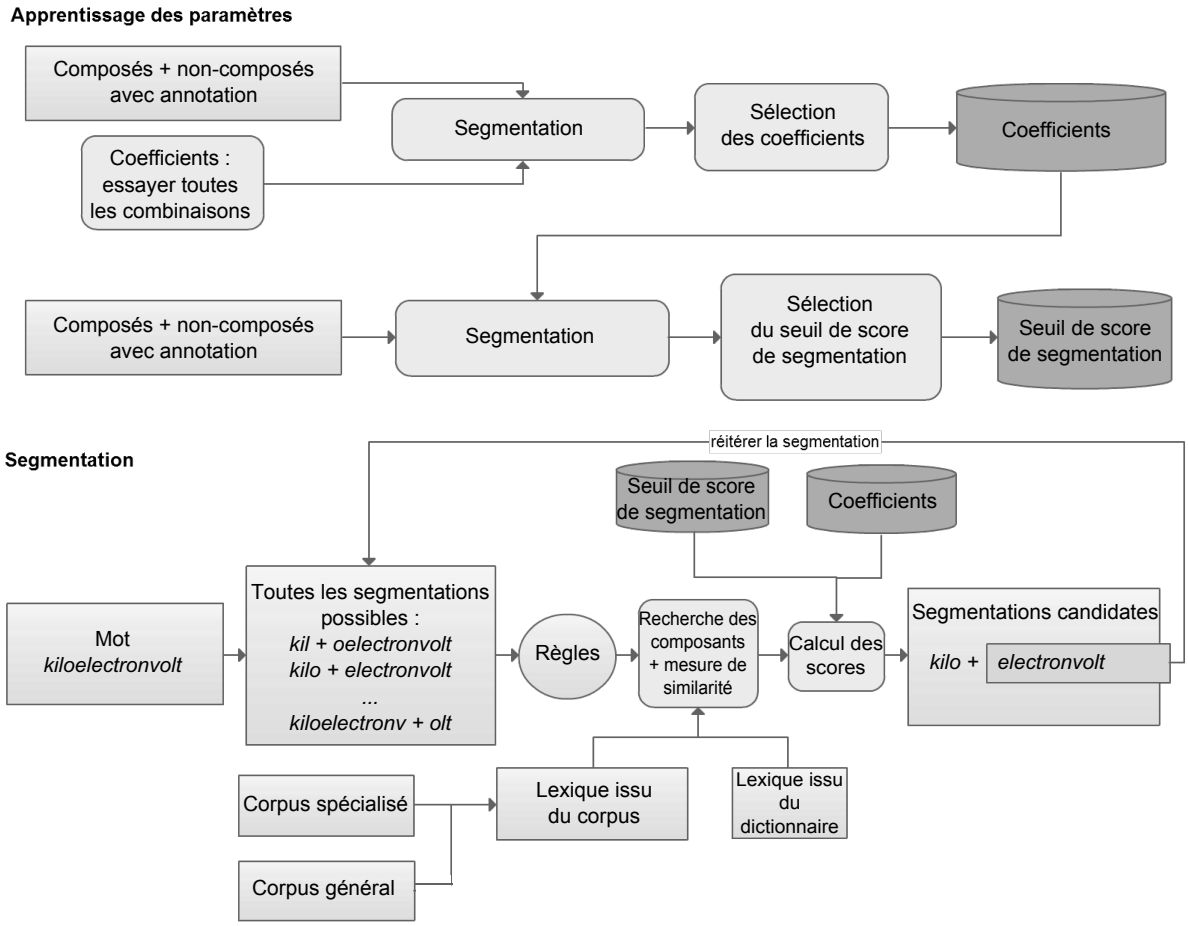

FIGURE 6.1 – Segmentation et apprentissage des paramètres

Pour réduire le volume des données, CompoST manipule des chaînes de caractères contenant seulement des lettres en minuscules, ce qui explique les minuscules au début des mots allemands dans les exemples.

Si des règles de transformation des composants en lexèmes indépendants sont disponibles pour la langue donnée, elles sont appliquées aux composants candidats. Dans l'exemple allemand cité, la règle *"s"* → *""* peut être appliquée : *Traktions* → *Traktion*. S'il n'y a pas de règles disponibles ou si elles ne couvrent pas toutes les transformations, les lemmes potentiels sont proposés en utilisant une mesure de similarité afin de choisir les lemmes « les plus proches » pour traiter l'allomorphie.

Pour chaque segmentation candidate, les lemmes pour les parties gauche et droite sont recherchés dans deux lexiques : le premier étant construit à partir du corpus monolingue spécialisé, et le deuxième combinant des entrées d'un dictionnaire général avec une liste d'éléments néoclassiques et de préfixes (pour les détails sur les lexiques construits, cf. annexe B.2). Ensuite, nous calculons le score de cette segmentation candidate. Le calcul du score est détaillé dans la partie 6.1.4.

La partie droite est soumise à la segmentation récursivement jusqu'à un certain niveau qui correspond à un nombre maximum de composants. Sur la base d'une étude des données, nous avons autorisé un maximum de 4 composants pour la langue allemande et de 3 pour les autres langues dans nos expériences.

RU килоэлектронвольт 'kiloélectronvolt' :
kiloelektronvol't → kilo + elektronvol't
elektronvol't → elektron + vol't

Enfin, l'algorithme propose les segmentations candidates classées par score décroissant (le nombre maximal de segmentations générées est un paramètre). Par exemple, pour DE *Traktionsbatterie* 'batterie

de traction' le résultat affiché est le suivant :

traktion + batterie 0.85
trakt + ion + batterie 0.82
transaktions + batterie 0.78
traut + ion + batterie 0.77
traktat + ion + batterie 0.75

Cinq segmentations ont été proposées (le nombre maximal de segmentations était fixé à 5). La segmentation correcte est *Traktion* + *Batterie*, et celle-ci obtient le meilleur score d'après CompoST.

6.1.3 Sélection des lemmes candidats pour un composant

Pour un composant, nous prenons pour lemmes candidats le composant-même, les chaînes de caractères générées à partir de ce composant par des règles, et certains lemmes issus des lexiques (annexe B.2) qui sont « similaires » au composant. Pour limiter le nombre de lemmes candidats, nous prenons seulement les lemmes ayant 3 caractères initiaux communs avec le composant, et nous définissons un seuil minimal de similarité entre le composant et son lemme égal à 0,7 (cette valeur étant déterminée de manière expérimentale).

Pour juger à quel point les chaînes de caractères sont proches, différentes mesures de similarité peuvent être utilisées (cf. description détaillée des mesures existantes Frunza et Inkpen (2009)). Dans nos travaux, nous utilisons « la distance d'édition normalisée » basée sur la distance de Levenshtein, qui est calculée ainsi :

$$sim(comp, lemma) = 1 - \frac{nbEditOper}{max(length(comp), length(lemma))} \tag{6.1}$$

où $nbEditOper$ est le nombre minimal d'opérations d'édition (substitution, suppression, insertion) nécessaires pour transformer un composant en un lemme. Cette mesure a été choisie parmi d'autres (préfixe commun le plus long, bigrammes partagés) suite aux expériences préliminaires sur un nombre réduit de données.

6.1.4 Score de segmentation

Nous partageons le principe théorique défendu par Benveniste (1974) selon lequel il y a dans le processus de composition toujours deux et seulement deux constituants (même si chacun entre eux peut être à son tour un composé). En accord avec notre *hypothèse 1* (section 5.1.7), nous calculons **le score d'une segmentation candidate** SC à chaque niveau de décomposition à partir du score de ses parties gauche et droite :

$$ScoreSegm(SC) = \begin{cases} \frac{Score(compA)+Score(compB)}{2} & \text{si correspondance exacte} \\ \frac{Score(compA)+Score(compB)}{nbComp} & \text{sinon} \end{cases}$$

où $nbComp$ est le nombre de composants à ce niveau de segmentation, et « correspondance exacte » signifie que tous les composants ont été trouvés en l'état dans un des lexiques utilisés. Le score d'une partie qui est elle-même décomposable est calculé comme le score d'une segmentation.

Par exemple pour un composé RU ветроколеса *vetrokolesa*$_{N.PL}$ 'roues éoliennes'
la segmentation correcte est : vetrokolesa = vetro:veter 'vent' + kolesa:koleso 'roue',

$$ScoreSegm(vetro + kolesa) = \frac{Score(veter) + Score(koleso)}{2} \tag{6.2}$$

la segmentation incorrecte est : vetrokolesa = vetro:veter 'vent' + ko:ko$_{PREF}$ 'co-' + lesa:les$_{N.PL}$ 'forêt'.

$$ScoreSegm(vetro + ko + lesa) = \frac{Score(veter) + \frac{Score(ko)+Score(les)}{2}}{3} \tag{6.3}$$

Toutefois, si tous les composants d'une segmentation correspondent à 100 % à leurs lemmes (similarité égale 1), il est fort probable que cette segmentation soit correcte. Il paraît donc logique de ne pas pénaliser de telles segmentations. Pour calculer le score de segmentation dans le cas de correspondance exacte, nous divisons la somme des scores des composants par 2, et non par $nbComp$.

$$ScoreSegm(kilo + elektron + volt) = \frac{Score(kilo) + \frac{Score(elektron)+Score(volt)}{2}}{2} \tag{6.4}$$

Le score d'un composant qui n'est pas décomposable est calculé par interpolation linéaire :

$$Score(C) = \alpha.sim(C, LC) + \beta.inDico(LC) + \gamma.inCorpus(LC) + \delta.dataCorpus(LC) \tag{6.5}$$

où C est un composant, LC est un lemme candidat pour ce composant, $sim(C, LC)$ signifie la similarité entre le composant et son lemme candidat (de 0 à 1), $inDico(LC)$ et $inCorpus(LC)$ sont des valeurs attestant respectivement de la présence ou l'absence du lemme candidat dans le lexique issu du dictionnaire et dans celui issu du corpus (0 ou 1), et $dataCorpus(LC)$ (entre 0 et 1 exclus) est une mesure de représentativité du lemme candidat dans le corpus (cf. section 6.1.5). Quant aux éléments néoclassiques NC du lexique qui n'apparaissent pas dans le corpus isolément, on leur assigne la valeur $inCorpus(NC)$ = 1, sinon le score serait pénalisant pour les composés néoclassiques.

Les coefficients α, β, γ et δ sont des paramètres qui doivent être appris individuellement pour chaque langue en utilisant une petite quantité de données d'apprentissage (une centaine de composés par langue plus un certain nombre de non-composés proportionnel au taux de non-composés dans cette langue). La somme des coefficients α, β, γ et δ est égale à 1.

6.1.5 Adaptation au domaine

Puisque notre travail porte sur la segmentation des termes relevant d'un domaine de spécialité, nous avons calculé la valeur $dataCorpus$ (cf. équation 6.5) non comme la fréquence d'un composant, mais comme sa spécificité au sein du domaine en question. Rappelons la définition de la spécificité (« *weirdness ratio* », Ahmad et al. (1992)) introduite dans la section 3.1 :

$$\textit{Spécificité(X)} = \frac{Freq_{LSP}(X)}{Freq_{LGP}(X)} \tag{6.6}$$

où $Freq_{LSP}(X)$ renvoie à la fréquence relative du mot X dans un corpus spécialisé, et $Freq_{LGP}(X)$ renvoie à la fréquence relative de ce mot dans un corpus général.

Nous exploitons cette mesure pour calculer les scores des composants candidats proposés par le système (cf. équation 6.5). Cependant, sa valeur peut être supérieure à 1, et interpolation linéaire oblige, tous les composants de la somme doivent être compris entre 0 et 1. Nous normalisons donc la spécificité d'un lemme en le divisant par la spécificité maximale existante pour ce corpus spécialisé $\textit{Spécificité}_{max}$:

$$DSpec(X) = \frac{\textit{Spécificité(X)}}{\textit{Spécificité}_{max}} \tag{6.7}$$

La spécificité aide à désambiguïser les variantes de la segmentation. Ainsi, pour le terme *Traktionsbatterie* du domaine de l'énergie éolienne, nous voudrions classer l'analyse *Traktion* + *Batterie* à un rang plus

élevé que *Trakt* + *Ion* + *Batterie*. Nous nous appuyons pour cela sur le fait que le mot *Traktion* est plus spécifique pour ce corpus que les mots *Trakt* et *Ion*.

Il reste possible avec la même méthode de traiter les composés de la langue générale. Dans ce cas, la valeur $dataCorpus$ d'un lemme est calculée comme sa fréquence relative dans un corpus général à la place de la spécificité. Nous allons tester par la suite l'impact de la spécificité sur la qualité de segmentation des termes.

6.1.6 Adaptation à la langue

Presque toutes les méthodes de segmentation de l'état de l'art tentent de s'adapter à la langue traitée. *L'anti-dictionnaire* (« *stop-list* »), i.e. une liste des lemmes des mots qui ne forment jamais de composés, est souvent utilisé pour éviter de proposer de fausses segmentations contenant ces mots. Les connaissances d'une langue peuvent être introduites sous forme de liste d'éléments de liaison qui s'insèrent entre les composants (Koehn et Knight, 2003), de règles de transformation des composants qui comportent l'insertion et l'omission de certains caractères (Ott, 2005), ou de grammaire contenant des règles et des restrictions sur l'usage de celles-ci (Schmid et al., 2004). Stymne (2008) et Weller et Heid (2012) ajoutent une restriction sur la catégorie grammaticale de la tête du composé : elle doit appartenir à la même catégorie grammaticale que le composé entier. Dans les langues germaniques auxquelles ces auteurs appliquent leurs méthodes, la tête du composé est toujours le composant le plus à droite.

La restriction sur la catégorie grammaticale du composant droit améliore la qualité de segmentation pour certaines langues, mais elle n'est pas universelle. Notamment dans les langues romanes le composant gauche peut être la tête du composé (Melloni et Bisetto, 2010; Scalise et Fabregas, 2010), e.g. FR *bateau-mouche* ou ES *camposanto* 'cimetière', lit. 'champ + saint'. Pour pouvoir traiter un spectre plus large de langues, nous n'avons pas intégré cette restriction dans CompoST.

En revanche, nous avons prévu l'utilisation optionnelle d'un anti-dictionnaire (cf. annexe B.2) et de règles qui transforment les composants non-autonomes en mots autonomes (en forme lemmatisée). Cette « normalisation » des composants aide à les reconnaître dans un dictionnaire ou dans un lexique construit à partir d'un corpus.

Normalisation des composants

Les règles de transformation des composants sont spécifiques à une langue donnée. Nous distinguons les règles qui s'appliquent au composant gauche et au composant droit. Les règles qui s'appliquent au composant gauche servent en premier lieu à traiter l'allomorphie entre les composants et les unités lexicales autonomes. Le composant droit doit parfois être modifié pour corriger la lemmatisation ou pour capturer la dérivation. Dans ce travail, nous allons appliquer CompoST à quatre langues typologiquement différentes. Nous décrivons ci-dessous les règles formulées pour chaque langue.

Transformations du composant gauche. L'allomorphie entre les composants et les unités lexicales autonomes est plus ou moins fréquente et régulière selon la langue.

L'allemand est une langue très compositionnelle avec allomorphie des composants fréquente et relativement régulière. La composition allemande est bien décrite dans la littérature. Pour transformer un composant en son lemme, dans la plupart des cas on doit soit supprimer l'élément de liaison "s" ("es") ou "n" ("en"), soit restaurer la désinence « perdue » dans le processus de composition. Les règles peuvent être formalisées par une liste quasiment exhaustive (cf. le tableau 6.1) basé sur (Langer, 1998). Les transformations

TABLE 6.1 – Règles de transformation du composant gauche en allemand basées sur (Langer 1998)

No.	Transformation	Exemple
1	"s" → ""	Staatsfeind : Staats → Staat
2	"n" → ""	Soziologenkongreß : Soziologen → Soziologe
3	"en" → ""	Straußenei : Straußen → Strauß
4	"er" → ""	Geisterstunde : Geister → Geist
5	"es" → ""	Geisteshaltung : Geistes → Geist
6	"en" → "us"	Aphorismenshatz : Aphorismen → Aphorismus
7	"en" → "um"	Museenverwaltung : Museen → Museum
8	"a" → "um"	Aphrodisiakaverkäufer : Aphrodisiaka → Aphrodisiakum
9	"en" → "a"	Madonnenkult : Madonnen → Madonna
10	"e" → ""	Hundehalter : Hunde → Hund
11	"en" → "on"	Stadienverbot : Stadien → Stadion
12	"a" → "on"	Pharmakaanalyse : Pharmaka → Pharmakon
13	"ien" → ""	Prinzipienreiter : Prinzipien → Prinzip
14	"i" → "e"	Carabinierischule : Carabinieri → Carabiniere
15	"" → "en"	Südwind : Süd → Süden
16	"" → "e"	Kirchhof : Kirch → Kirche

contenant un *umlaut* (e.g. DE *Gänseklein* 'abats d'oie' = *Gans* + *Klein*) ne font pas partie des règles utilisées dans nos expériences. Nous supposons que ces cas seront traités grâce à l'utilisation de la similarité de chaînes de caractères.

La composition en russe est moins régulière qu'en allemand, mais également productive, surtout dans les domaines spécialisés. Les voyelles "o" et "e" servent d'éléments de liaison. La désinence du composant gauche est souvent omise. La langue russe a un large système de flexion. En nous basant sur l'étude classique de la morphologie russe de Zalizniak (1977), nous avons établi une liste de règles pour restaurer les désinences des composants (cf. tableau 6.2). Cette liste n'est pas exhaustive, mais elle couvre les transformations les plus fréquentes. Ces règles ne couvrent pas les cas d'altération de la base, comme dans l'exemple :

RU ветрогенератор 'générateur éolien'
vetrogenerator = veter 'vent' + generator 'générateur';

Ces transformations, de même que l'umlaut en allemand, sont supposées être traitées grâce à l'utilisation de la mesure de similarité. Nous allons tester deux jeux de règles pour le russe : le premier jeu de règles « règles basiques » ne contient que deux règles traduisant le fait que les voyelles "o" et "e" servent d'éléments de liaison; le deuxième jeu de règles « règles élargies » contient toutes les règles présentées dans le tableau 6.2.

Il existe des études linguistiques consacrées à la composition en anglais (Bauer, 1983; Lieber, 2010), majoritairement à la composition syntagmatique. L'anglais produit aussi des composés morphologiques, mais généralement par simple concaténation des composants : *airfoil, workgroup, middleground, streamtube*. Par conséquent, nous n'avons pas introduit de règles pour le traitement d'allomorphie pour cette langue. Par contre, parmi les quasi-composés anglais, certains présentent un composant gauche sous forme fléchie, par exemple *lessons-learned* 'bonne pratique', 'leçons tirées' ou *smaller-scale* dans l'expression *smaller-scale wind turbine* 'éolienne de plus petite taille'. Pour traiter ces cas, nous avons appliqué au composant gauche en anglais les mêmes règles qu'au composant droit.

En français la composition n'est pas productive, à l'exception de la composition néoclassique et d'un type particulier de composition native (*lave-vaisselle*, *porte-manteau*) qui s'est avéré peu présent dans les deux corpus spécialisés en question. Les composants natifs en français sont rarement soumis à modification, nous n'avons donc pas introduit de règles de transformation du composant gauche pour cette langue.

TABLE 6.2 – Règles de transformation du composant gauche en russe

No.	Contexte gauche	Transformation	Exemple
1	-	"о" → ""	kapitalo → kapital
2	-	"е" → ""	sredne → sredn(ij)
3	-	"о" → "а"	vodo → voda
4	-	"е" → "я"	zemle → zemlâ
5	"ж/ш/щ/ч/ц"	"е" → "а"	tysâče → tysâča
6	-	"е" → "ь"	žizne → žizn’
7	-	"о" → "ый"	krupno → krupnyj
8	-	"о" → "ой"	krivo → krivoj
9	-	"е" → "ий"	obše → obšij
10	"к/г"	"о" → "ий"	vysoko → vysokij

Transformations du composant droit. En plus des règles qui décrivent les transformations aux frontières des composants (i.e. du composant gauche), nous avons introduit des règles de transformation du composant droit. Ces règles sont utiles pour le traitement de la flexion et de la dérivation :

1. La lemmatisation effectuée par un outil automatique statistique échoue souvent pour les termes composés (qui ne sont pas présents dans les données d’apprentissage), et nous obtenons ainsi des formes fléchies dans la liste des mots à analyser, par exemple *endométriaux*. Les règles introduites aident à lemmatiser les composés.
2. Les adjectifs dénominaux sont rarement inclus dans les dictionnaires, particulièrement ceux dérivés de noms très spécialisés. On peut analyser les composés contenant de tels adjectifs en transformant ces derniers vers les noms dont ils sont dérivés :

   RU adjectif паротурбинный ’à turbine à la vapeur’
   paro.turbinnyj = paro:par ’vapeur’ + turbinnyj:turbinnyj$_{ADJ}$ (n’est pas dans le dictionnaire)
   turbinnyj → turbina$_N$ ’turbine’ (dans le dictionnaire).

Le tableau 6.3 récapitule les règles de transformation du composant droit que nous avons définies pour chaque langue. Pour les cas où les règles ne suffisent pas (*centrique* → *centr*, et non *centre*), la similarité de chaînes peut contribuer à retrouver le lemme.

6.1.7 Adaptation au type de composé

Notre objectif est de traiter de la même manière tous les types de phénomènes que nous avons décrits (composés natifs, néoclassiques, empruntés, mots préfixés et quasi-composés). Les composés natifs et quasi-composés ne nécessitent pas de ressources complémentaires, les lemmes de leurs composants appartenant au dictionnaire ou au corpus. Pour le traitement des composés empruntés, nous nous reposons également sur le corpus en supposant que leurs composants empruntés y apparaissent indépendamment (ce qui est vrai pour une partie des empruntés). Les composants néoclassiques, ainsi que les préfixes, n’apparaissent pas isolément dans les textes, à quelques rares exceptions près. Pour les identifier, nous pouvons utiliser des listes de ces éléments pour chaque langue.

Les composants néoclassiques ont une origine latine ou grecque, par conséquent ils ont une forme proche, y compris dans des langues éloignées. De plus, ce phénomène est bien décrit dans la littérature (Amiot et Dal, 2008; Namer, 2009; Béchade, 1992). Cela permet d’obtenir une liste d’éléments néoclassiques pour une nouvelle langue relativement facilement. Nous avons complété ces listes d’éléments néoclassiques par certains préfixes d’origine latine, grecque ou autre qui ont des équivalents dans de nombreuses langues (*co-*, *pré-*, *post-*, *trans-*, etc.). Nous allons nous référer à ces ressources comme **NCP-listes**. Les détails concernant la construction des NCP-listes et leur taille sont donnés dans l’annexe B.2.

TABLE 6.3 – Règles de transformation du composant droit

Transformation	Exemple
DE	
"n" → ""	Achsen → Achse
"en" → ""	Türen → Tür
"innen" → "in"	Schülerinnen → Schülerin
"e" → ""	Betriebe → Betrieb
"se" → "s"	Ergebnisse → Ergebnis
"ungen" → "ung"	Wicklungen → Wicklung
"s" → ""	Betreibers → Betreiber
"es" → ""	umweltamtes → umweltamt
EN	
"s" → ""	tubes → tube
"ed" → ""	coated → coat
"ed" → "e"	based → base
"ing" → ""	erecting → erect
"ing" → "e"	erasing → erase
"ies" → "y"	species → specy
"er" → ""	smaller → small
"ier" → "y"	easier → easy
"est" → ""	smallest → small
"iest" → "y"	easiest → easy
"al" → ""	technical → technic
FR	
"s" → ""	variables → variable
"e" → ""	paramétrée → paramétré
"ique" → ""	centrique → centr
"iques" → ""	magnétiques → magnétique
"aux" → "al"	originaux → original
"aux" → "au"	réseaux → réseau
"aux" → "ail"	coraux → corail
RU	
"ный" → ""	ètažnyj → ètaž
"ной" → ""	skorostnoj → skorost
"овой" → ""	silovoj → sil
"овый" → ""	volnovyj → voln
"евой" → ""	poršnevoj → poršn
"евый" → ""	višnevyj → višn
"ской" → ""	morskoj → mor
"ский" → ""	kavkazskij → kavkaz
"чатый" → ""	stupenčatyj → stupen
"ический" → ""	cilindričeskij → cilindr
"яный" → ""	glinânyj → glin

6.1.8 Apprentissage des paramètres

Jusqu'à présent, nous n'avons pas expliqué comment les paramètres du système sont fixés. Ils sont définis suite à l'entraînement du système sur un ensemble de lemmes annotés préalablement avec leurs segmentations. Cet ensemble contient des lemmes des composés ainsi que des non-composés. La phase d'apprentissage consiste en deux étapes : la définition des coefficients pour le calcul du score d'un composant, et la définition d'un seuil pour le score d'une segmentation qui sert à décider s'il s'agit d'un composé ou non.

Choix des coefficients. Afin de définir les coefficients α, β, γ et δ optimaux (équation 6.5), nous calculons les scores de segmentations pour les termes annotés avec toutes les valeurs possibles de chaque coefficient entre 0 et 1 avec un pas de 0,1 de manière à ce que la somme des coefficients soit égale à 1. Nous évaluons les résultats en termes de précision et de rappel (en comparant avec l'annotation correcte fournie) et nous retenons la combinaison qui produit les meilleurs résultats. A cette étape, les paramètres produisant la meilleure précision sont généralement les mêmes que ceux qui donnent le meilleur rappel. Plus de détails sur le calcul de la précision et du rappel sont exposés dans la section 6.2.2.

Identification d'un composé. Nous ré-appliquons le même algorithme avec les coefficients sélectionnés sur le même jeu de données annotées avec l'objectif de déterminer un seuil optimal de score de segmentation à partir duquel le mot donné est probablement un composé. Nous essayons toutes les valeurs de seuil entre 0 et 1 avec un pas de 0,05, et nous calculons le rappel et la précision pour le Top 1 et le Top 5 des segmentations candidates les mieux classées par le système et avec des scores supérieurs au seuil courant. Nous pouvons ensuite choisir le seuil que nous considérons optimal.

Nous supposons que les paramètres optimaux varient en fonction de la langue mais aussi des connaissances externes utilisées, notamment des règles, car la similarité des chaînes de caractères (coefficient α) peut être plus ou moins importante selon les langues, pour lesquelles des règles plus ou moins complètes sont disponibles, voire absentes. Nous allons donc définir les paramètres pour chaque langue.

Une fois que tous les paramètres sont sélectionnés, nous pouvons appliquer CompoST sur les données de test.

Bilan

Pour résumer, nous avons présenté un système multilingue permettant d'identifier et de segmenter des mots composés. Le système combine l'utilisation du corpus, du dictionnaire et de la mesure de similarité de chaînes de caractères. Optionnellement, il peut intégrer des règles de transformation des composants et des ressources lexicales complémentaires spécifiques à une langue. Le système est adaptable au domaine de spécialité grâce à l'utilisation de la spécificité des unités lexicales.

6.2 Évaluation intrinsèque des résultats de la segmentation

Nous effectuons une évaluation intrinsèque de notre système sur quatre langues et deux domaines de spécialité (énergie éolienne et cancer du sein). CompoST prévoit plusieurs options de segmentation. Nous testons différentes options pour évaluer l'impact de chaque composante du système, et nous analysons les résultats. Des extraits des résultats de la segmentation sont présentés dans l'annexe D.

6.2.1 Liste de référence de termes composés et non-composés

Pour évaluer le système, ainsi que pour déterminer les paramètres adéquats pour chaque langue, nous avons besoin d'un ensemble de lemmes, annotés en tant que composés ou non, et pour les composés, annotés avec leurs composants. Nous n'avons pas trouvé de telles listes pour les quatre langues de nos expériences. La campagne d'évaluation d'apprentissage de la morphologie Morpho Challenge (cf. section 5.2.5) fournit des données annotées pour l'anglais et l'allemand. Mais dans ces jeux de données, les mots composés ne sont pas distingués formellement des mots simples. Pour nos expériences, il serait possible d'en annoter manuellement une partie comme composés ou non, et de modifier l'annotation de manière à ignorer les affixes flexionnels et dérivationnels. Cela permettrait de disposer de données pour deux langues sur quatre, mais il resterait tout de même à en trouver pour le français et le russe. De plus, les données de Morpho Challenge relèvent de la langue générale, et non d'un domaine de spécialité. Pour ces raisons, nous n'adoptons pas ces données pour nos expériences.

Nous avons choisi de construire des listes de référence à partir des corpus spécialisés que nous disposons. Pour les deux domaines et pour chaque langue, nous avons extrait du corpus monolingue la liste des mots respectant les critères suivants :

1. L'absence de traduction pour ces mots dans le dictionnaire bilingue utilisé ;
2. L'appartenance à une des deux catégories - nom ou adjectif (d'après une étude multilingue de Scalise et Vogel (2010), 80 % de composés appartiennent à ces catégories). Pour l'étiquetage des catégories grammaticales, TreeTagger [2] a été utilisé ;
3. Une fréquence supérieure à 2 ou à 5 en fonction du corpus (2 pour le domaine de l'énergie éolienne et 5 pour le cancer du sein : le nombre de mots inconnus du dictionnaire avec la fréquence supérieure à 5 était très peu élevé pour le corpus de l'énergie éolienne, nous avons donc réduit le minimum à 2) ;
4. Une longueur minimale de 6 caractères qui correspond à un composé de deux parties de 3 caractères chacune (restriction définie auparavant sur la longueur du composant).

Ces critères servent à éliminer des mots ayant une probabilité très faible d'être composés et à diminuer le nombre de candidats à segmenter.

Les lexiques ainsi obtenus ont été ensuite triés dans un ordre aléatoire et annotés manuellement jusqu'à ce que nous obtenions entre 200 et 300 composés. Chaque composé a été annoté avec sa segmentation correcte (ou les segmentations correctes) et une des catégories que nous avions introduites : natif, néoclassique, emprunt, quasi-composé ou mot préfixé. Les autres mots ont été annotés comme non-composés. Nous avons réalisé l'annotation pour le russe. Les listes de référence pour d'autres langues ont été annotées par nos collègues. Chacune des listes EN et FR du domaine *énergie éolienne* a été annotée par deux annotateurs afin de pouvoir mesurer l'accord entre annotateurs.

Notre hypothèse était que les paramètres définis pendant la phase d'apprentissage dépendent en premier lieu de la langue traitée et non du domaine de spécialité (ou tout du moins dans une moindre mesure). Ainsi, pour chaque langue nous avons choisi un domaine de spécialité sur lequel entraîner le système. Il s'agit de *l'énergie éolienne* pour l'anglais et le russe, et du *cancer du sein* pour l'allemand et le français. Pour le domaine choisi, les données annotées ont été divisées en deux parties : *le jeu d'entraînement* utilisé pour l'apprentissage des paramètres du système et *le jeu de test* utilisé pour l'évaluation des résultats. Le tableau 6.4 récapitule la taille des listes de référence pour chaque corpus, ainsi que le taux de composés. Le tableau 6.5 montre la distribution des types de composés dans le jeu de test pour chaque langue et domaine.

On peut remarquer que les listes pour l'allemand contiennent le pourcentage le plus élevé de composés parmi les langues analysées, ce qui n'est pas surprenant. Les corpus du domaine du *cancer du sein*

2. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

TABLE 6.4 – Taille du lexique annoté et taux de composés

Langue et domaine	Nb. de lemmes	Nb. de composés	Taux de composés
	Jeu d'entraînement		
DE Cancer du sein	157	109	70 %
EN Energie éolienne	404	106	27 %
FR Cancer du sein	245	93	38 %
RU Energie éolienne	485	116	24 %
	Jeu de test		
DE Energie éolienne	162	103	64 %
DE Cancer du sein	150	109	73 %
EN Energie éolienne	401	100	25 %
EN Cancer du sein	503	201	40 %
FR Energie éolienne	413	107	26 %
FR Cancer du sein	340	160	47 %
RU Energie éolienne	700	170	24 %

TABLE 6.5 – Distribution des types de composés dans le jeu de test. NV signifie composés natifs, NC - composés néoclassiques, PR - mots préfixés, QC - quasi-composés, LN - composés empruntés (« *loan* »).

Langue et domaine	NV	NC	PR	QC	LN
DE Energie éolienne	89	2	4	0	8
DE Cancer du sein	73	32	3	0	1
EN Energie éolienne	55	14	19	12	0
EN Cancer du sein	40	94	40	27	0
FR Energie éolienne	12	44	40	11	0
FR Cancer du sein	11	95	51	3	0
RU Energie éolienne	98	53	16	0	3

contiennent plus de composés que les corpus sur l'*énergie éolienne*. La raison principale en est le fort pourcentage de termes néoclassiques dans les corpus du *cancer du sein*, caractéristiques des domaines médicaux. Les quasi-composés sont productifs dans les langues analytiques et ne le sont pas dans les langues synthétiques. Les emprunts sont apparus dans les corpus allemands et russes, probablement parce que la terminologie de ces langues présente moins de ressemblances avec la terminologie anglaise que celle du français.

Accord entre annotateurs

L'annotation manuelle pose toujours la question de l'objectivité. On constate souvent des divergences entre les annotations faites par des personnes différentes. Pour évaluer l'importance de l'erreur potentiellement introduite, et pour révéler les points problématiques d'une annotation, il est d'usage de calculer un *accord inter-annotateurs*. Dans ce but, des métriques variées ont été proposées dans la littérature, en fonction des données évaluées. Nous avons choisi le test du « kappa » introduit par Jacob Cohen (Carletta, 1996), conçu pour calculer l'accord entre deux annotateurs qui classifient N unités en M catégories exclusives, ce qui correspond à notre cas. Les unités sont des termes (composés ou non), et les catégories sont les suivantes : natif, néoclassique, emprunté, quasi-composé, préfixé et non-composé.

Le coefficient kappa montre la convergence des annotateurs en prenant en compte la probabilité d'un accord aléatoire. Il est calculé ainsi :

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{6.8}$$

où $P(A)$ est l'accord observé entre les annotateurs et $P(E)$ est la probabilité d'un accord aléatoire.

L'accord observé $P(A)$ est égal au nombre de cas dans lesquels les observateurs ont pris la même décision (par exemple, accord sur le fait que tel mot est néoclassique, ou que tel autre n'est pas un composé, etc.) divisé par le nombre total de mots analysés. La probabilité d'un accord aléatoire est calculée comme la somme des probabilités d'un accord aléatoire pour chaque catégorie. Pour une catégorie, la probabilité que les deux annotateurs prennent la même décision est égale au produit des probabilités indépendantes de cette décision pour le premier et le deuxième annotateurs, divisé par le nombre total de mots analysés. La probabilité d'une décision prise par un annotateur est calculée comme le nombre de fois où il a assigné cette catégorie à un mot, divisé par le nombre total de mots analysés.

L'accord a été évalué sur les termes du corpus de l'énergie éolienne pour deux langues : l'anglais et le français. Le coefficient obtenu est de 0,89 pour la liste anglaise et de 0,93 pour la liste française, ce qui entre dans la catégorie « accord presque parfait » selon l'échelle de Landis et Koch (1977) (entre 0,8 et 1). Les annotateurs n'étaient pas d'accord sur certains mots à mi-chemin vers la lexicalisation comme EN *outboard* (préfixé ou non-composé). Le point le plus controversé était la distinction entre des composés natifs et des quasi-composés, par exemple EN *case-control*, *variable-speed* 'vitesse variable', *twenty-year* qui sont reliés par un tiret quand ils font partie d'un composé syntagmatique plus large (*case-control study* 'étude de cas-témoin'). Pour la version de liste utilisée dans les expériences, nous avons choisi une étiquette pour chaque cas (ainsi les mots énumérés étaient marqués comme quasi-composés).

6.2.2 Mesures de l'évaluation

Les mesures habituelles pour la tâche de segmentation ont été exploitées pour l'évaluation : la précision, le rappel, la F-mesure et l'exactitude (cf. section 5.2.5).

Pour chaque lemme de la liste de référence, l'analyse produite par CompoST a été comparée avec l'annotation de référence. Rappelons que CompoST produit plusieurs segmentations ordonnées par pertinence : pour ce travail, le nombre maximal a été défini à 5. Toutes les mesures ont été calculées pour le Top 1 et pour le Top 5. Dans le premier cas, l'unité candidate pour l'évaluation est seulement la segmentation la

mieux classée. Dans le deuxième cas, l'unité candidate est l'ensemble des cinq propositions, lesquelles sont comparées avec la référence, et si parmi elles il y en a au moins une de correcte, l'analyse est considérée correcte.

Ainsi, la précision pour le Top 5 est calculée comme le ratio entre le nombre de mots qui ont une segmentation correcte parmi les 5 propositions du système, et le nombre de mots segmentés :

$$P_{TopN} = \frac{\textit{nb unités candidates ayant au moins une segmentation correcte dans le Top N}}{\textit{nb unités candidates segmentées}} \tag{6.9}$$

Le rappel pour le Top 5 est calculé comme le ratio entre le nombre de mots qui ont une segmentation correcte parmi les 5 propositions du système, et le nombre de composés présents dans la liste de référence :

$$R_{TopN} = \frac{\textit{nb unités candidates ayant au moins une segmentation correcte dans le Top N}}{\textit{nb composés dans la liste de référence}} \tag{6.10}$$

La F-mesure est calculée comme la moyenne harmonique entre la précision et le rappel :

$$F_{TopN} = \frac{2 \times P_{TopN} \times R_{TopN}}{P_{TopN} + R_{TopN}} \tag{6.11}$$

L'exactitude (« *accuracy* ») est calculée ainsi :

$$A_{TopN} = \frac{\textit{nb unités candidates analysées correctement}}{\textit{nb total de lemmes dans la liste de référence}} \tag{6.12}$$

6.2.3 Configurations du système

CompoST permet plusieurs options de segmentation : on peut intégrer des ressources lexicales supplémentaires (NCP-liste, anti-dictionnaire) et/ou des règles de transformation des composants. L'adaptation au domaine est théoriquement assurée par l'utilisation de la spécificité des unités lexicales, mais la fréquence dans le corpus de la langue générale peut être utilisée à la place de la spécificité.

Afin d'évaluer l'impact de chaque option, nous avons testé plusieurs configurations du système :

1. BASE : la configuration de base sans utilisation de ressources lexicales, ni de règles;
2. BASE + NCP + STOP : la configuration de base enrichie par des ressources lexicales ;
3. BASE + RÈGLES : la configuration de base enrichie par des règles, mais sans ressources lexicales. Notons que pour le russe nous avons deux jeux de règles, basique (RÈGLES1) et élargi (RÈGLES2), ce qui introduit une configuration supplémentaire ;
4. BASE + NCP + STOP + RÈGLES : la configuration complète enrichie à la fois par des ressources lexicales et des règles.

Pour évaluer l'impact de la spécificité, nous avons également testé une autre variante de la configuration complète, celle avec l'utilisation de la fréquence dans le corpus de la langue générale au lieu de la spécificité (« Fréquence générale » vs. « Spécificité »).

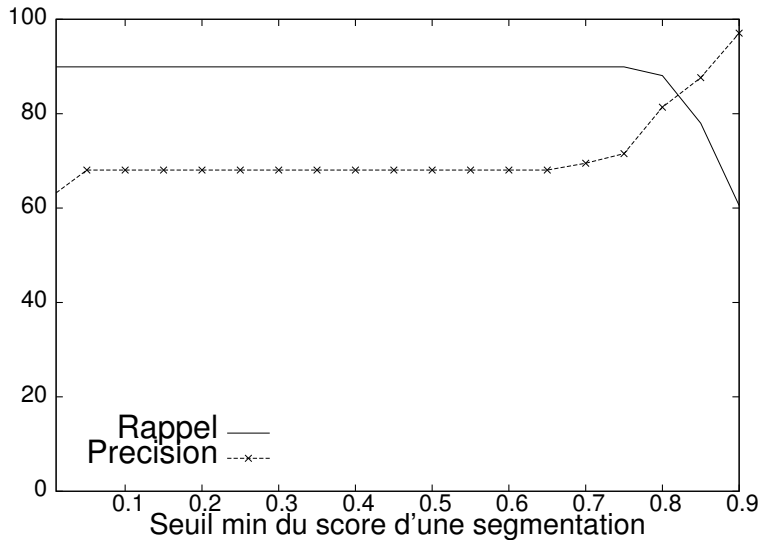

FIGURE 6.2 – Précision/rappel en fonction du seuil minimal du score d'une segmentation (DE Top 1)

6.2.4 Paramètres retenus

Nous supposons que :

1. Les paramètres du système doivent être appris pour chaque configuration séparément, car ils sont sensibles à la présence/absence de règles et aux ressources utilisées (ces ressources étant spécifiques à la langue) ;
2. Les paramètres sont relativement peu sensibles au domaine de spécialité.

Pour chaque langue, nous avons d'abord défini les meilleurs paramètres pour les configurations énumérées sur le jeu de données d'entraînement dans un domaine, nous avons appliqué ensuite ces paramètres sur le jeu de test du même domaine afin de choisir la meilleure configuration, et nous avons enfin testé cette configuration avec les paramètres retenus sur le jeu de tests du deuxième domaine.

Coefficients. Les expériences ont confirmé que les coefficients optimaux diffèrent selon la langue et la configuration des ressources (pour les valeurs des paramètres retenues pour chaque configuration, cf. annexe C). Parfois, nous avons été confrontée à la situation suivante : pour la même configuration, plusieurs combinaisons de coefficients produisent exactement les mêmes résultats sur les données d'entraînement. Un choix est toutefois nécessaire. Nous admettons également que les paramètres obtenus peuvent ne pas être optimaux car il est possible que les meilleurs paramètres puissent être détectés avec un « pas » plus petit, par exemple, 0,001 au lieu de 0,1. Cependant, cela augmenterait considérablement le temps de traitement, sans garantir pour autant l'obtention des paramètres les plus performants. Dans de tels cas, nous avons fixé des valeurs de paramètres sans perdre de vue qu'un certain biais est possible, et nous allons vérifier sur les données de test si les paramètres retenus aboutissent tout de même à un résultat acceptable.

Seuil optimal du score d'une segmentation. Le seuil optimal du score d'une segmentation sert à identifier si un mot est composé ou non. En outre, il permet de choisir si l'on souhaite privilégier le rappel ou la précision selon l'application finale. Par exemple, on privilégiera la précision pour l'acquisition automatique du lexique spécialisé, et le rappel pour la recherche d'information ou la traduction supervisée. Pour illustrer le choix du seuil, nous présentons sous la forme d'une courbe la précision et le rappel pour le Top 1 en DE (jeu d'entraînement) en fonction du seuil défini (figure 6.2). Nous retenons pour les tests deux configurations du seuil : l'une permettant d'optimiser le rappel, et l'autre permettant d'optimiser la précision. Dans

TABLE 6.6 – Qualité de la segmentation en fonction des connaissances utilisées. R (%) signifie rappel, P (%) - précision, F (%) - F-mesure, A (%) - exactitude (accuracy). Les meilleurs résultats sont mis en gras.

Configuration	Rappel optimal		Précision optimale	
	Top 1	Top 5	Top 1	Top 5
	R P F A	R P F A	R P F A	R P F A
DE Cancer du sein				
BASE	73 61 66 65	84 70 76 73	72 81 76 77	80 89 84 83
BASE+NCP+STOP	70 64 67 69	84 77 80 79	67 75 71 73	75 85 80 79
BASE+RÈGLES	78 75 76 77	87 84 85 83	**77** 79 78 77	**84** 87 85 83
BASE+NCP+STOP+RÈGLES	**79 77 78 80**	**90 88 89 88**	75 **84 79 81**	83 **92 87 86**
EN Énergie éolienne				
BASE	86 57 69 82	88 59 71 82	74 **74** 74 88	75 75 75 88
BASE+NCP+STOP	89 **59 71 83**	92 **61** 73 **84**	82 73 77 89	83 74 78 89
BASE+RÈGLES	83 54 65 80	87 57 69 81	83 70 76 88	87 74 80 89
BASE+NCP+STOP+RÈGLES	**90** 58 **71 83**	**94 61 74 84**	**87 74 80 91**	**91 78 84 92**
FR Cancer du sein				
BASE	43 46 44 59	48 51 49 61	30 69 42 63	30 69 42 63
BASE+NCP+STOP	78 **69 73 78**	83 **73 78 80**	**65** 83 73 80	**66** 84 74 80
BASE+RÈGLES	44 44 44 58	48 48 48 60	39 73 51 67	40 74 52 68
BASE+NCP+STOP+RÈGLES	**80** 60 69 71	**87** 66 75 75	**65 91 76 82**	65 **91 76 82**
RU Énergie éolienne				
BASE	34 20 25 63	49 29 36 67	29 32 30 76	42 46 44 79
BASE+NCP+STOP	64 35 45 70	72 39 51 72	**60** 53 56 83	**65** 58 61 84
BASE+RÈGLES1	36 21 27 63	59 34 43 68	33 33 33 76	52 53 52 81
BASE+RÈGLES2	45 39 42 77	58 50 54 80	39 50 44 81	50 65 57 83
BASE+NCP+STOP+RÈGLES2	**77 48 59 78**	**86 54 66 81**	52 **81 63 86**	52 **82 64 86**

l'exemple DE ci-dessus, nous avons choisi la valeur 0,8 pour le cas où le rappel est privilégié, et 0,85 pour la précision.

6.2.5 Impact des ressources lexicales supplémentaires

Les résultats obtenus avec des configurations variées sont présentés dans le tableau 6.6.

Globalement nous pouvons constater une certaine accumulation : les résultats BASE+NCP+STOP et BASE+RÈGLES sont meilleurs que la BASE, et les résultats de la configuration complète sont meilleurs que ceux de BASE+NCP+STOP et BASE+RÈGLES prises séparément. Toutefois nous remarquons quelques exceptions.

Pour DE, la configuration BASE+NCP+STOP donne des résultats inférieurs à la configuration BASE pour toutes les mesures d'évaluation avec le seuil optimisé pour la précision, et en termes de rappel pour le Top 1 avec le seuil optimisé pour le rappel. Une conclusion hâtive serait que l'usage de ressources lexicales supplémentaires nuit aux résultats pour l'allemand. La raison est cependant plus complexe. D'une part, le dictionnaire utilisé pour DE contient beaucoup de termes médicaux, voire de racines néoclassiques (e.g., *kryo-*), ce qui explique l'obtention de bons résultats sans NCP-liste. D'autre part, notre méthode ne prend pas en compte la position initiale ou finale de l'élément. Avec l'utilisation des préfixes, des segmentations candidates erronées contenant une sous-chaîne semblable à un préfixe peuvent apparaître, comme *in* dans l'exemple ci-dessous :

DE Tyrosinkinasehemmer 'inhibiteur de tyrosine-kinase'
Segmentation correcte (BASE): tyrosin + kinase + hemmer;
Segmentation incorrecte (BASE+NCP+STOP): tyros + in + kinase + hemmer;

Ce défaut se propage sur la configuration BASE+NCP+STOP+RÈGLES qui produit donc un gain assez modeste par rapport à la BASE (3-4 points) et même une perte de rappel par rapport à la BASE+RÈGLES avec le seuil optimisé pour la précision. Pour comparer, avec le seuil optimisé pour le rappel, le gain de la BASE+NCP+STOP+RÈGLES par rapport à la BASE varie entre 6 et 18 points.

Pour EN et FR, la configuration BASE+RÈGLES avec le seuil optimisé pour le rappel produit des résultats inférieurs à ceux de BASE+NCP+STOP et même de BASE. L'utilisation de règles introduit dans certains cas des faux positifs :

EN laboratories → laboratory (RÈGLE : "ies" → "y") = lab + oratory;
FR histopathologique → histopatholog (RÈGLE : "ique" → "") = histopathologie + que.

L'usage de l'anti-dictionnaire peut remédier à ce défaut (FR *que* est filtré), mais pas dans tous les cas. En conséquence la précision dans la configuration BASE+NCP+STOP+RÈGLES avec le seuil optimisé pour le rappel est inférieure ou égale à celle de BASE+NCP+STOP.

Pour RU, la configuration BASE+NCP+STOP+RÈGLES avec le seuil optimisé pour la précision perd en rappel par rapport à la BASE+NCP+STOP (8 points pour le Top 1 et 13 points pour le Top 5). En revanche, elle gagne largement en précision (28 et 24 points respectivement). Le jeu de règles élargi perd légèrement en rappel pour le Top 5 pour les deux seuils (1-2 points), mais gagne largement en précision (12-16 points).

Malgré ces observations, dans la plupart des cas la configuration la plus complète a produit les meilleurs résultats. Par ailleurs, nous pouvons remarquer que pour EN, FR et RU l'usage de ressources lexicales complémentaires apporte un gain plus important que les règles, tandis que pour DE les règles améliorent considérablement les résultats. Cela s'explique par le fait que, d'un côté, la formation des composés en DE est assez régulière, et de l'autre côté, le dictionnaire utilisé, comme nous l'avons dit plus haut, contient déjà un nombre important de radicaux néoclassiques.

6.2.6 Impact de la spécificité

Nous avons comparé l'utilisation de la spécificité avec l'utilisation de la fréquence dans un corpus général (cf. tableau 6.7). La spécificité permet d'obtenir des résultats supérieurs pour DE, EN et FR. Pour RU les résultats sont mitigés : parfois le rappel ou la précision sont meilleurs avec la spécificité, parfois avec la fréquence générale. Les valeurs de F-mesure et d'exactitude se trouvent par conséquent très proches pour deux configurations, à l'exception du cas de la précision optimale (Top 1), dans lequel nous constatons un écart de 15 points en précision en faveur de l'utilisation de la spécificité qui résulte en un écart de 5 points pour la F-mesure. La spécificité semble moins déterminante que la fréquence générale pour le russe. La raison est à chercher dans les ressources : pour le russe, nous avons exploité une liste de fréquences publiquement accessible qui était compilée à partir du *Corpus national du russe*, un corpus très complet et équilibré (cf. annexe B.1), tandis que pour les autres langues nous avons compilé la liste équivalente à partir des corpus bruts journalistiques. De plus, le corpus russe du domaine de l'énergie éolienne compilé à partir du web peut se révéler moins spécialisé que les corpus éoliens d'autres langues, et les termes extraits sont par conséquent moins spécialisés. A notre connaissance, il n'existe pas de mesure permettant d'estimer la spécialisation d'un corpus par rapport à un domaine.

TABLE 6.7 – Impact de la spécificité sur la segmentation. R (%) signifie rappel, P (%) - précision, F (%) - F-mesure, A (%) - exactitude (accuracy). Les meilleurs résultats sont mis en gras.

Configuration	Rappel optimal				Précision optimale			
	Top 1				Top 5			
	R	P	F	A	R	P	F	A
DE Cancer du sein								
Spécificité	**79**	**77**	**78**	**80**	**90**	**88**	**89**	**88**
Fréquence générale	77	65	70	69	89	75	81	77
EN Énergie éolienne								
Spécificité	**90**	**58**	**71**	**83**	**94**	**61**	**74**	**84**
Fréquence générale	87	49	63	76	92	51	66	77
FR Cancer du sein								
Spécificité	**80**	**60**	**69**	**71**	**87**	**66**	**75**	**75**
Fréquence générale	78	55	65	66	81	58	68	68
RU Énergie éolienne								
Spécificité	**77**	48	**59**	78	**86**	54	**66**	81
Fréquence générale	60	**52**	56	**80**	71	**62**	**66**	**83**

Configuration	Précision optimale							
	Top 1				Top 5			
	R	P	F	A	R	P	F	A
DE Cancer du sein								
Spécificité	**75**	**84**	**79**	**81**	**83**	**92**	**87**	**86**
Fréquence générale	73	80	76	79	**83**	91	**87**	**86**
EN Énergie éolienne								
Spécificité	**87**	**74**	**80**	**91**	**91**	**78**	**84**	**92**
Fréquence générale	84	67	75	87	88	70	78	88
FR Cancer du sein								
Spécificité	65	**91**	**76**	**82**	65	**91**	**76**	**82**
Fréquence générale	**71**	78	74	81	**71**	78	74	81
RU Énergie éolienne								
Spécificité	**52**	**81**	**63**	**86**	52	**82**	64	**86**
Fréquence générale	**52**	66	58	84	**58**	74	**65**	**86**

L'utilisation de la spécificité reste plutôt positive pour le traitement des composés issus des domaines de spécialité. Y compris pour RU, elle fournit un ordonnancement efficace. Ainsi, pour le composé RU ветротурбина *vetroturbina* 'turbine éolienne' avec la spécificité, le classement proposé par CompoST est :

vetroturbina :
veter turbina 0.81
veto turbina 0.8
vetka turbina 0.8

Avec l'utilisation de la fréquence générale, les candidats ont le même score. Le lemme *veter* 'vent' est obtenu à partir du composant *vetro* transformé en *vetr* par la règle "o" $\rightarrow$ "", la similarité $sim(veter, vetr)$ étant égale à 0,8; le lemme *veto* 'veto' est obtenu à partir du composant *vetro*, la similarité $sim(veto, vetro)$ étant égale à 0,8; et le lemme *vetka* 'branche' est obtenu à partir du composant *vetro* transformé en *vetra* par la règle "o" $\rightarrow$ "a", la similarité $sim(vetra, vetka)$ étant aussi égale à 0,8. Ainsi, les couples de chaînes $(veter, vetr)$, $(veto, vetro)$ et $(vetra, vetka)$ ont la même valeur de similarité, ce qui explique les scores égaux de trois segmentations candidates obtenus avec l'utilisation de la fréquence générale. En revanche, *veter* 'vent' a une spécificité beaucoup plus élevée que les deux autres dans le domaine de l'énergie éolienne, ce qui permet d'ordonner les segmentations candidates. L'utilité de la spécificité a déjà été démontrée pour l'extraction terminologique dans plusieurs langues (Ahmad et al., 1992; Daille et Blancafort, 2013).

6.2.7 F-mesure vs. exactitude

Nous avons présenté dans la section 5.2.5 deux mesures permettant de caractériser la performance d'un système de segmentation : la F-mesure et l'exactitude. Un examen des résultats obtenus montre que dans toutes nos expériences, à une exception près (spécificité vs. fréquence générale, RU, rappel optimal, Top 1) les meilleurs choix d'après les deux mesures coïncident. Quand les valeurs du rappel et de la précision sont proches (cas DE), la F-mesure et l'exactitude sont également proches. En revanche, quand il y a un écart important entre le rappel et la précision, la valeur d'exactitude est bien plus élevée que celle de la F-mesure et subit des variations moindres. L'exactitude est très sensible à la quantité de composés dans les données analysées. Si l'exactitude était calculée pour un système qui ne segmente aucun mot, elle serait égale au pourcentage de non-composés dans les données, c'est-à-dire à 76 % pour RU (en guise de

comparaison : dans notre meilleure configuration de la segmentation, l'exactitude est égale à 78 %). Par contre pour DE (énergie éolienne) elle serait égale à 36 %. Nous en concluons que la F-mesure est plus adaptée à l'évaluation de la segmentation et nous allons par la suite nous limiter à son calcul.

6.2.8 Analyse quantitative des résultats

La configuration la plus complète du système, i.e. intégrant les ressources lexicales et les règles linguistiques et utilisant la spécificité, s'est globalement avérée la meilleure. Nous l'avons appliqué sur le jeu de test du deuxième domaine qui n'avait pas servi à l'entraînement : *énergie éolienne* pour DE et FR, *cancer du sein* pour EN. Les paramètres retenus pour chaque langue avec cette configuration sont présentés dans le tableau 6.8. Les résultats pour les deux domaines sont présentés dans les tableaux 6.9 (avec un seuil optimisé pour le rappel) et 6.10 (avec un seuil optimisé pour la précision). Pour chaque langue, le domaine qui a servi pour l'entraînement est indiqué en gris.

TABLE 6.8 – Paramètres retenus

Langue	Domaine	Coefficients (α β γ δ)	Seuil	
			rappel optimal	précision optimale
DE	Cancer du sein	0,5 0,3 0,1 0,1	0,8	0,85
EN	Énergie éolienne	0,7 0,1 0,1 0,1	0,8	0,85
FR	Cancer du sein	0,5 0,1 0,1 0,3	0,6	0,7
RU	Énergie éolienne	0,3 0,1 0,4 0,2	0,7	0,8

TABLE 6.9 – Évaluation de la segmentation avec le seuil optimisé pour le rappel.
R (%) signifie rappel, P (%) - précision, F (%) - F-mesure.
En gris - le domaine qui a servi pour l'apprentissage des paramètres.

Langue	Énergie éolienne						Cancer du sein					
	Top 1			Top 5			Top 1			Top 5		
	R	P	F	R	P	F	R	P	F	R	P	F
DE	79	68	73	87	75	81	79	77	78	90	88	89
EN	90	58	71	94	61	74	84	60	70	86	62	72
FR	85	51	64	90	54	68	80	60	69	87	66	75
RU	77	48	59	86	54	66						

Avec le seuil du score d'une segmentation optimisé pour le meilleur rappel (tableau 6.9), le rappel est compris entre 77 % et 90 % pour le Top 1, et entre 86 % et 94 % pour le Top 5. La précision est comprise entre 48 % et 77 % pour le Top 1, et entre 54 % et 88 % pour le Top 5. La F-mesure est comprise entre 59 % et 78 % pour le Top 1, et entre 66 % et 89 % pour le Top 5. Avec le seuil optimisé pour la meilleure précision (tableau 6.10), le rappel varie entre 52 % et 87 % pour le Top 1, et entre 52 % et 91 % pour le Top 5. La précision varie entre 74 % et 93 % pour le Top 1 et entre 76 % et 94 % pour le Top 5. La F-mesure est comprise entre 63 % et 82 % pour le Top 1, et entre 64 % et 87 % pour le Top 5.

Langue. Avec un seuil optimisé pour le rappel, le rappel est généralement meilleur que la précision (ce qui n'est pas surprenant étant donné le choix du seuil). Seulement pour l'allemand la précision est relativement haute et côtoie le rappel. La raison est la suivante : en allemand le taux des composés dans les listes de référence est nettement plus élevé que dans les autres langues (cf. le tableau 6.4), par conséquent le nombre

TABLE 6.10 – Évaluation de la segmentation avec le seuil optimisé pour la précision.
R (%) signifie rappel, P (%) - précision, F (%) - F-mesure.
En gris - le domaine qui a servi pour l'apprentissage des paramètres.
Les résultats qui surpassent ceux obtenus avec la méthode (Koehn et Knight 2003) sont mis en gras.

Langue	Énergie éolienne						Cancer du sein					
	Top 1			Top 5			Top 1			Top 5		
	R	P	F	R	P	F	R	P	F	R	P	F
DE	**78**	77	**77**	**83**	83	**83**	**75**	84	**79**	**83**	**92**	**87**
EN	**87**	74	**80**	**91**	78	**84**	**77**	76	**76**	**77**	76	76
FR	**73**	**93**	**82**	74	**94**	**83**	**65**	**91**	**76**	65	**91**	76
RU	**52**	**81**	**63**	52	**82**	64						

de non-composés parmi les candidats qui peuvent introduire une fausse segmentation est moindre. Inversement, le taux de composés dans le domaine de l'énergie éolienne en RU, FR et EN est très bas (autour de 25 %), et la précision est faible.

Pour les deux configuration du seuil, les résultats les moins bons ont été obtenus pour le russe. Cela peut s'expliquer par l'abondance des transformations des composants dans cette langue. Nous allons montrer dans la section 6.3 que le russe représente un défi également pour d'autres systèmes de segmentation.

Type de composé. La performance du traitement varie aussi en fonction du type de composé. Les composés néoclassiques sont généralement bien identifiés et segmentés sous réserve que leurs composants soient recensés dans nos NCP-listes. Les composés natifs sont les plus difficiles à segmenter à cause des transformations des composants. La capacité à segmenter un mot préfixé ou un composé emprunté dépend de l'existence de ses composants respectivement dans la NCP-liste ou le corpus. Les quasi-composés sont faciles à segmenter car ils contiennent un trait d'union : *fréquence-puissance* = *fréquence* + *puissance*.

Nous avons remarqué la présence de composés dits parasynthétiques dans le lexique extrait du corpus RU. Ils ont été annotés comme des composés natifs. De telles formations sont parfois analysées correctement par notre système grâce aux règles de transformations et à la similarité de chaînes de caractères :

ВЫСОКОВОЛЬТНЫЙ$_{ADJ}$ 'haut voltage'
vysokovol'tnyj → vysoko + vol'tnyj
vysoko → vysokij 'haut' (RÈGLE : "ko" → "kij")
vol'tnyj → vol't 'volt' (RÈGLE : "nyj" → "").

Mais dans d'autres cas, la segmentation échoue. L'analyse des composés parasynthétiques sort du cadre purement compositionnel et nécessite l'inventaire complet des règles morphologiques pour chaque langue.

Domaine de spécialité. Nous avons supposé que les paramètres du système dépendent de la langue et donc de la configuration des ressources utilisées spécifiques à la langue, et non du domaine (ou non significativement). Nous avons appliqué la meilleure configuration et les paramètres retenus sur les données du domaine qui n'avait pas servi à l'apprentissage des paramètres (*énergie éolienne* pour DE et FR, *cancer du sein* pour EN). Nous observons qu'avec le seuil optimisé pour le rappel les résultats sont cependant meilleurs pour le domaine d'entraînement. La même tendance est visible avec le seuil optimisé pour la précision, mais moins marquée : pour FR, les résultats sont meilleurs pour l'énergie éolienne, même si l'apprentissage a été effectué sur les termes du corpus du cancer du sein. Les valeurs de la F-mesure obtenues sur un nouveau domaine sont pour la plupart du même ordre de grandeur que celles du domaine d'apprentissage. Nous constatons tout de même un écart supérieur ou égal à 9 points en précision pour le FR avec le seuil optimisé pour le rappel (tableau 6.9), et en rappel pour EN avec le seuil optimisé pour la précision

TABLE 6.11 – Évaluation de la segmentation après la correction de la lemmatisation

Domaine	Top 1			Top 5		
	R	P	F	R	P	F
DE Énergie éolienne	79	74	76	87	82	84
DE Cancer du sein	84	86	85	93	95	94

(tableau 6.10). Nous concluons que l'entraînement sur le même domaine de spécialité que les termes traités est plus efficace. Néanmoins l'utilisation des paramètres appris sur un domaine pour un autre est tout à fait possible et permet d'obtenir des résultats « décents ».

6.2.9 Analyse qualitative des erreurs

L'analyse des résultats de la segmentation nous a conduite à distinguer deux types d'erreurs : les erreurs dues au pré-traitement et les erreurs intrinsèques résultant de la méthode appliquée.

Erreurs du pré-traitement

Correction de la lemmatisation. La lemmatisation correcte est très importante pour la bonne segmentation et la reconnaissance des composés. Dans nos expériences, la lemmatisation a été effectuée par un outil probabiliste entraîné sur les données de la langue générale, de ce fait certains mots spécialisés, composés ainsi que non-composés, n'ont pas été lemmatisés correctement (ou tout simplement pas lemmatisés). Les règles de transformation du composant droit que nous avons introduites pour traiter ce genre de cas aident à segmenter correctement des composés non-lemmatisés, mais elles ne résolvent pas le problème pour les mots non-composés (cf. EN *laboratories* → *laboratory* = *lab* + *oratory*). La correction de la lemmatisation devrait avoir lieu avant la vérification dans le dictionnaire. Cela permettrait d'identifier certains mots comme faisant partie du lexique et donc ne pas les segmenter, ce qui à son tour diminuerait le nombre de fausses segmentations et améliorerait la précision.

Pour prouver cette hypothèse, nous avons mené une expérience préliminaire sur DE visant la correction de la lemmatisation : les désinences nominales et adjectivales *(e)s, (e)n, er, em, (e)ns, e* ont été omises à la fin des mots à condition qu'un mot sans une de ces terminaisons apparaisse soit dans le dictionnaire, soit dans le corpus. C'est une variante « brute » de la correction de lemmatisation pour DE, proposée par Anita Gojun [3]. La correction de la lemmatisation entraîne une amélioration importante de la précision (de 7 à 9 points) sans nuire au rappel (cf. le tableau 6.11 vs. le tableau 6.9).

Entités nommées. Certains cas d'erreurs dits « faux positifs » sont dus à la segmentation des entités nommées, e.g. DE *Denmark* a été segmenté en *den* + *mark*. Pour le RU, cette erreur représente 10 % des faux positifs (13 sur 124), et parmi ces 13 unités, seulement 5 ont été étiquetées comme nom propre par TreeTagger.

Erreurs intrinsèques

Ordre des composants. Parmi les erreurs dues à notre méthode, nous pouvons constater celles qui sont liées à l'ordre des composants. Comme nous l'avons mentionné plus haut, la position des composants n'a

3. http://www.ims.uni-stuttgart.de/institut/mitarbeiter/gojunaa/software.html

pas été prise en compte, ce qui a mené à une segmentation incorrecte dans certains cas (cf. DE *Tyrosinkinasehemmer*). Cependant, dans certaines langues, le composant droit peut être préfixé sans être séparé par le tiret du composant gauche :

RU инсулинонезависимый 'insulino-indépendant'
insulinonezavisimyj = insulin 'insuline' + o + ne$_{PREF}$ + zavisimyj 'dépendant'.

En plus des règles spécifiques à la langue, des règles précisant une certaine combinatoire peuvent être envisagées, par exemple : un préfixe peut apparaître soit au début du mot, soit après le tiret, soit après un autre préfixe, soit après un élément de liaison.

Mesure de similarité. Enfin, notre méthode recourt à l'utilisation de la mesure de similarité. Par conséquent, certains affixes ou combinaisons d'affixes peuvent se confondre avec des unités lexicales autonomes, par exemple :

RU керосиновый 'à pétrole', adjectif non-composé
Segmentation incorrecte: kerosinovyj = kerosin$_N$ 'pétrole' + novyj 'nouveau'

Pour contourner ce problème, un large inventaire d'affixes et de bases est nécessaire pour chaque langue.

En revanche, l'utilisation de la similarité permet de reconnaître et segmenter des composés natifs dans le cas où les transformations de leurs composants ne sont pas décrites par la liste de règles utilisées, e.g. RU *vetr* → *veter* 'vent' (cf. exemple RU *vetrogenerator*) ou DE *Verbände* → *Verband* (*Verbändevereinbarung* 'accord d'association' = *Verband* 'association' + *Vereinbarung* 'accord').

Nous avons décrit les expériences effectuées et évalué les résultats du point de vue quantitatif et qualitatif. Pour situer ces résultats par rapport à l'état de l'art sur la segmentation automatique, nous allons dans la partie suivante comparer CompoST avec trois autres méthodes.

6.3 Comparaison avec l'état de l'art

Afin de comparer les résultats de notre méthode avec les autres, deux méthodes parmi celles décrites dans la section 5.2 ont été choisies : Morfessor (Virpioja et al., 2013) et (Koehn et Knight, 2003). Les deux approches sont indépendantes de la langue, ce qui permet de les appliquer à nos données. La première est une méthode probabiliste et la deuxième est une méthode basée sur le corpus. Pour des exemples de segmentations produites par chaque approche, cf. tableau 6.12.

Pour se comparer à une méthode purement lexicale (i.e. qui découpe les mots à l'aide d'un lexique d'éléments), nous avons établi une équivalence entre l'évaluation faite dans le travail de (Delpech, 2013) utilisant la segmentation lexicale comme étape pour la traduction, et notre évaluation. Cela a été possible grâce à une intersection entre les jeux de test utilisés dans les deux travaux.

6.3.1 Comparaison avec une méthode probabiliste

Morfessor est un système non-supervisé conçu pour l'analyse morphologique, mais il peut également être utilisé pour la segmentation des mots composés. Nous avons utilisé Morfessor 2.0[4] entraîné sur nos corpus spécialisés et en ajoutant les NCP-listes. Puisque nous avons rajouté aux données d'entraînement les NCP-listes qui contiennent seulement les formes de composants néoclassiques et de préfixes, et ne contiennent pas leur fréquence dans le corpus, nous avons choisi l'option d'entraînement « *type-based training* » : chaque forme fléchie n'est comptée qu'une seule fois sans prendre en compte sa fréquence. Ensuite

4. http://www.cis.hut.fi/projects/morpho/morfessor2.shtml, l'entraînement a été mené avec l'algorithme récursif.

TABLE 6.12 – Exemples de segmentation. *Analyse correcte* renvoie à la segmentation correcte en composants et non à l'analyse morphologique complète.

	Mot	Traduction FR	Analyse correcte	CompoST	Morfessor	Koehn and Knight (2003)
DE	Asynchron-maschinen	machines asynchrones	asynchron maschine	asynchron maschine	Asynchron-maschine n	Asynchron-maschine
	ergebend	résultant	ergebend	ergeben end	ergeben d	ergebend
EN	impermissible	inadmissible	im permissible	impermissible	impermiss ible	impermissible
	anthropogenic		anthropo genic	anthropo genic	an thro po genic	anthropogenic
FR	mini-réseaux		mini réseau	mini réseau	mini réseaux	mini réseau
	paramétrées		paramétrées	para métrie	paramétr ées	paramétrées
RU	ветро-турбины	turbines éoliennes	ветер турбина	ветер турбина	ветротурбин ы	ветротурбины
	окислитель	oxydant	окислитель	окислитель	окис ли тель	окислитель

TABLE 6.13 – Évaluation de la segmentation faite par Morfessor sur le domaine de l'*énergie éolienne*

	No. mots	No. mots segmentés	No. composés	No. composés bien segmentés	R Top 5 (%)
DE	162	125	103	50	49
EN	401	250	100	84	83
FR	413	300	107	77	72
RU	700	629	170	31	18

nous avons réalisé la segmentation de nos données de test du domaine *énergie éolienne*. Les résultats sont présentés dans le tableau 6.13.

Morfessor a été conçu pour segmenter les mots morphologiquement complexes en morphèmes, ce qui explique le grand nombre de mots segmentés. Dans ce travail, nous ne cherchons pas à évaluer la segmentation en morphèmes, mais seulement celle des composés en composants. Dans notre évaluation nous ignorons donc les découpages erronés des affixes flexionnels et dérivationnels. Pour les mots composés, il est aisé de juger si la segmentation proposée est correcte. Mais pour les mots non-composés qui ont été segmentés d'une manière erronée il est difficile de statuer sur une erreur de segmentation en morphèmes ou en composants, cf. *negatif* $\rightarrow$ *ne* + *g* + *atif* ou *circonférentiel* $\rightarrow$ *circon* + *f* + *érentiel*. Face à cette difficulté, nous avons choisi de ne pas prendre en compte les non-composés, et nous n'avons donc évalué que le seul rappel.

Comme nous l'avons déjà évoqué, Morfessor a été conçu pour les langues ayant une morphologie concaténative, par conséquent il n'effectue pas la lemmatisation des composants. Par exemple le terme FR *mini-réseaux* a été segmenté en *mini* + *réseaux* au lieu de *mini* + *réseau*. Cette analyse pourrait être considérée comme correcte si le but était seulement l'établissement des frontières entre les composants. Nous avons évalué de tels cas comme des erreurs car notre but est de trouver les lemmes des composants. Cela explique pourquoi le rappel est relativement haut pour l'anglais, légèrement plus bas pour le français et assez bas pour les langues flexionnelles comme l'allemand et le russe. L'évaluation faite sur les lemmes favorise Com-

poST qui normalise les composants à l'inverse de Morfessor qui ne fait que segmenter. Il existe une autre méthode de la même famille, Allomorfessor (Virpioja et al., 2009), prévue pour proposer une solution pour le traitement de l'allomorphie. Mais actuellement cette méthode n'est pas disponible au téléchargement.

Quant aux autres erreurs, Morfessor entraîné avec l'option « *type-based training* » a un comportement assez conservatif, c'est-à-dire qu'il a tendance à ne pas segmenter certains mots qui devraient l'être, e.g. *aérogénérateur*. Toutefois, la sur-segmentation est aussi possible : EN *anthropogenic* = *an* + *thro* + *po* + *genic*.

Même avec le seuil optimisé pour la précision (cf. tableau 6.10), la configuration de CompoST mettant en œuvre des règles surpasse Morfessor en rappel, avec une fourchette de 2 à 34 % selon les langues, sachant que l'écart de 34 % a été constaté pour les deux langues ayant une morphologie riche : l'allemand et le russe. Si on compare avec une configuration de CompoST n'utilisant pas de règles (cf. tableau 6.6, BASE+NCP+STOP), le rappel en EN est de 83 %, i.e. égal au rappel de Morfessor, et en RU de 65 % contre 18 % pour Morfessor.

6.3.2 Comparaison avec une méthode basée sur un corpus

L'approche proposée par Koehn et Knight (2003) est devenue l'état de l'art de la segmentation des mots composés avec l'utilisation de corpus. Nous l'avons appliquée à nos données en rajoutant les règles de transformation et en exploitant les mêmes ressources (corpus spécialisé, anti-dictionnaire, NCP-liste) que dans l'évaluation de CompoST.

Pour un mot donné, cette méthode produit une ou plusieurs analyses candidates y compris le mot initial non-segmenté (cf. section 5.2.3). Les candidats sont classés selon un score égal à la moyenne géométrique des fréquences des composants dans le corpus, et pour un mot non-segmenté, à sa fréquence.

Afin de pouvoir exploiter la NCP-liste avec cette méthode basée sur les fréquences des mots dans le corpus, nous avons artificiellement affecté la fréquence de 1 aux éléments appartenant à la NCP-liste. Sinon le système n'aurait pas traité correctement les composés néoclassiques et préfixés contenant des composants non-autonomes. Cependant, la valeur de 1 n'est pas forcément judicieuse car elle peut désavantager les segmentations contenant des éléments NCP. Une méthode alternative pour réduire ce biais serait d'affecter aux éléments de la NCP-liste leur fréquence réelle dans le corpus en tant que sous-chaînes d'autres mots. Cependant cette approche poserait le problème opposé, à savoir une fréquence surélevée des éléments NCP car ils seraient parfois comptés par erreur : e.g. *pré-* apparaît comme sous-chaîne des mots *préserver, prétention, préventif*, etc.

Contrairement à notre méthode qui qualifie chaque mot soit de composé, soit de non-composé, avec la méthode (Koehn et Knight, 2003) le mot non-segmenté est un candidat parmi les autres, c'est-à-dire qu'il peut apparaître au premier, deuxième ou n-ème rang dans le classement des résultats. Cela complique la comparaison des résultats pour le Top 5. Nous avons décidé d'évaluer l'analyse d'un composé comme réussie si une segmentation correcte apparaît parmi les 5 premiers candidats, même si le mot non-segmenté est classé au premier rang. Les résultats en terme de rappel, précision et F-mesure sont présentés dans le tableau 6.14.

Cette méthode, couplée aux règles de transformation, donne une bonne précision car elle produit peu de segmentations erronées. Dans le même temps, cette approche est connue pour laisser non-segmentés les composés dont la fréquence dans le corpus est plus élevée que la fréquence de leurs composants. Ainsi, pour DE *Asynchronmaschine* 'machine asynchrone' le meilleur candidat est *asynchronmaschine* et la segmentation *asynchron* + *maschine* vient en seconde position.

Puisque cette approche privilégie la précision par rapport au rappel, il est justifié de comparer les résultats à ceux obtenus avec CompoST en utilisant les paramètres optimisés pour la meilleure précision (cf. tableau 6.10, les nombres en gras correspondent aux cas dans lesquels notre méthode s'est révélée plus

performante que (Koehn et Knight, 2003). En comparant la précision, les approches sont à peu près équivalentes car pour 7 expériences sur 14 (Koehn et Knight, 2003) s'avère plus précis, CompoST étant plus performant dans les 7 autres. La méthode (Koehn et Knight, 2003) est moins précise pour le français et le russe, mais plus précise que notre méthode pour l'allemand et l'anglais (avec un écart allant jusqu'à 13 points). Pour ces mêmes expériences, le rappel de notre méthode est plus élevé, jusqu'à 34 points (EN Cancer du sein). Au global, notre méthode donne le meilleur rappel dans 11 expériences sur 14 et la meilleure F-mesure dans 12 expériences.

Pour conclure la comparaison de ces trois approches, Morfessor s'est montré le moins performant parmi les trois pour la segmentation des mots composés car il n'est pas adapté pour les langues flexionnelles. La méthode (Koehn et Knight, 2003) produit une précision meilleure que le rappel. Notre méthode peut être paramétrée pour privilégier soit la précision, soit le rappel. Avec les paramètres optimisant la précision, notre méthode est en moyenne aussi précise que (Koehn et Knight, 2003) tout en permettant de segmenter un plus grand nombre de composés, surtout pour le Top 1.

6.3.3 Comparaison avec une méthode lexicale

Le travail de (Delpech, 2013) est consacré à la traduction des termes construits morphologiquement. Pendant la première étape du traitement, elle effectue la segmentation des termes en morphèmes : préfixes, suffixes, radicaux néoclassiques et populaires (natifs). La décomposition des mots se fait à l'aide d'une méthode que nous qualifions de *lexicale* : les composants sont découpés s'ils sont présents dans un lexique. Le lexique utilisé comprend une liste de morphèmes liés (y compris des éléments néoclassiques) construite préalablement, ainsi qu'une liste d'unités lexicales autonomes combinant des entrées d'un dictionnaire de la langue générale, des entrées d'un dictionnaire de synonymes, et une liste des formes extraites du corpus. Si plusieurs segmentations sont possibles, une seule est utilisée pour la traduction, celle ayant le plus grand nombre de composants. Delpech (2013) accomplit la traduction de l'anglais (langue avec une composition plutôt concaténative) vers le français et l'allemand, elle ne traite donc pas les transformations des composants.

Cette segmentation, comme dans le cas de Morfessor, est une tâche différente de la segmentation en composants car, premièrement, son but est de découper les mots en sous-chaînes et non de trouver leurs lemmes et, deuxièmement, les sous-chaînes correspondent aux morphèmes et non aux composants (analyse plus « morcelée »). Néanmoins, il est possible d'établir une correspondance partielle entre les deux évaluations car (1) les termes traités sont issus du même corpus EN du cancer du sein, et (2) il s'agit de la langue anglaise pour laquelle les composants sont souvent identiques à leurs lemmes.

Afin de comparer les résultats, nous avons sélectionné l'intersection entre les termes morphologiquement complexes analysés par (Delpech, 2013) et notre jeu de données de test, soit 73 termes composés au total. Parmi eux, 71 termes étaient bien décomposés par la méthode lexicale utilisée par (Delpech, 2013), et

TABLE 6.14 – Évaluation de la segmentation faite par la méthode (Koehn and Knight 2003). R (%) signifie le rappel, P (%) - la précision, F (%) - la F-mesure.

Langue	Énergie éolienne						Cancer du sein					
	Top 1			Top 5			Top 1			Top 5		
	R	P	F	R	P	F	R	P	F	R	P	F
DE	60	86	71	67	87	76	61	89	72	66	91	77
EN	71	87	78	74	83	78	43	82	56	74	82	78
FR	65	83	73	76	84	80	22	90	35	68	87	76
RU	48	69	57	62	68	65						

70 termes ont été bien segmentés par notre méthode avec le seuil optimisé pour la précision. Les résultats sont très proches et dans le même temps, les erreurs ne sont pas les mêmes (nous avons constaté 9 analyses différentes). La méthode lexicale produit parfois de la sur-segmentation car c'est une analyse privilégiant le plus grand nombre de composants (*œstrogen-dependent* = *œstro* + *gen* + *de* + *pendent, relapse-free* = *re* + *lapse* + *free*), dans les cas où CompoST avec l'utilisation de la spécificité ne la commet pas. Les ressources lexicales utilisées dans deux évaluations se croisent (le lexique des mots du corpus, la table de morphèmes liés construite par (Delpech, 2013) que nous avons exploitée), mais elles ne sont pas identiques : (Delpech, 2013) utilise le dictionnaire de langue générale et le dictionnaire de synonymes intégrés dans l'analyseur linguistique XELDA [5]. La différence des ressources explique que parfois des termes segmentés dans le travail de Delpech (2013) ne le sont pas avec notre système (*pharmacokinetic*) ou à l'inverse (*cytokeratin*).

6.4 Bilan

Dans cette partie, nous avons présenté une méthode de reconnaissance et de segmentation des composés, et nous l'avons évaluée sur les termes de quatre langues typologiquement différentes. Les résultats sont comparables avec ceux obtenus avec d'autres méthodes d'état de l'art.

Contrairement à la méthode lexicale et la méthode probabiliste (Morfessor), CompoST effectue non seulement la segmentation en composants, mais aussi la recherche des lemmes des composants à l'aide de règles linguistiques. Il permet également d'utiliser la spécificité dans un corpus spécialisé. Par rapport à la méthode de Koehn et Knight (2003), basée sur la fréquence des mots dans le corpus et adaptable pour intégrer des règles linguistiques, CompoST exploite un dictionnaire et la similarité de chaînes de caractères. De plus, il permet de paramétrer le ratio entre la précision et le rappel pour l'adapter à l'application visée. Il faut aussi noter que c'est une méthode supervisée qui demande un certain nombre de données annotées que les méthodes citées ne demandent pas.

Nos expériences confirment que l'utilisation de la spécificité pour le traitement de termes est globalement bénéfique car elle permet de remonter des segmentations candidates très spécialisées au premier rang. L'utilisation de règles linguistiques et de ressources complémentaires (listes d'éléments néoclassiques et de préfixes, anti-dictionnaire) améliore généralement la qualité de segmentation. L'utilisation de la similarité des chaînes de caractères est plus controversée : elle introduit un certain nombre d'erreurs et diminue la précision; en revanche, elle augmente le rappel en permettant d'analyser des composés natifs avec les transformations de composants non-couvertes par les règles.

Nous avons délimité préalablement les phénomènes que nous visions : les composés natifs, néoclassiques, empruntés, ainsi que des mots préfixés et les quasi-composés. Le niveau de difficulté du traitement automatique n'est pas égal pour ces catégories. Le traitement des composés néoclassiques et préfixés avec notre méthode repose sur la présence de leurs éléments dans le lexique, par conséquent nous obtenons une précision plus élevée que le rappel. Pour les composés natifs, au contraire, le rappel est généralement plus haut que la précision. Les quasi-composés anglais et français sont souvent séparés par un trait d'union et donc faciles à segmenter. Nous avons également constaté que ce type de composition n'est pas productif dans des langues synthétiques.

Nos expériences soulignent une fois de plus l'importance du pré-traitement en TAL. Les résultats pourraient être améliorés en corrigeant en amont la lemmatisation. La reconnaissance des entités nommées est également utile pour les distinguer des mots composés qui devraient être segmentés.

Les résultats d'un système de segmentation des mots composés peuvent être utilisés dans de nombreuses applications du TAL. Nous présentons dans la partie suivante son utilisation pour la traduction compositionnelle et pour l'identification des variantes des termes composés.

5. http://www.temis.com

Traitement des termes composés

Nous abordons dans cette dernière partie de la thèse deux aspects de la construction de lexiques terminologiques qui nécessitent l'identification et la segmentation automatique des termes composés : l'alignement des termes composés de la langue source avec leurs équivalents de la langue cible, et la détection des variantes terminologiques (au sein de la même langue).

L'alignement des termes composés d'une langue où la composition morphologique est productive, avec les termes polylexicaux d'une langue moins compositionnelle requiert la segmentation des termes dans la langue source. Cette tâche de l'alignement lexicale se rapproche de la traduction des termes, nous allons donc l'étudier ici dans un contexte de la traduction compositionnelle. La traduction compositionnelle peut toutefois constituer une application en soi (et non dans le cadre de l'extraction terminologique), lorsque la segmentation des mots composés peut aussi enrichir la traduction automatique assistée par ordinateur ou la traduction automatique statistique. D'ailleurs c'est souvent dans ce dernier contexte que son utilité a été évaluée.

Au niveau monolingue, la segmentation peut être utile pour regrouper les termes composés avec leurs variantes polylexicales. Nous allons détecter automatiquement les variantes de ce type dans quatre langues abordées, et nous allons comparer la productivité de cette variation et les structures des variantes identifiées selon la langue.

7

Traduction compositionnelle

Les termes composés représentent un défi pour la traduction automatique et la traduction en général car la plupart d'entre eux ne sont pas recensés dans les dictionnaires bilingues non-spécialisés. Le fait qu'un composé morphologique se traduit souvent vers d'autres langues par un composé syntagmatique (polylexical) complique la tâche de la traduction automatique.

Les méthodes de traduction automatique statistique (TAS) se basent sur l'apprentissage à partir de grands ensembles de données (des phrases alignées avec leurs traductions). Cependant, ces données sont habituellement extraits de corpus généraux, et les termes composés morphologiques, relativement rares dans la langue générale, ne s'y trouvent pas, ou s'y trouvent avec une fréquence trop faible pour garantir une traduction correcte. Le repérage et la segmentation des termes composés sont par conséquent utiles pour TAS, ce qui a été démontré dans plusieurs travaux (Koehn et Knight, 2003; Dyer, 2009; Macherey et al., 2011). L'utilité de la segmentation pour la traduction compositionnelle a été également étudiée (Weller et Heid, 2012). Nous poursuivons cette deuxième piste, la traduction compositionnelle, car contrairement à la TAS, elle ne requiert pas de corpus parallèles, toujours difficiles à obtenir pour des domaines spécialisés.

L'objectif principal de ce chapitre est d'évaluer l'impact de la segmentation réalisée par notre système sur la qualité de la traduction compositionnelle des termes. L'objectif secondaire est de vérifier quelle configuration de notre système - à savoir celle qui optimise le rappel ou la précision - sera plus efficace pour cette tâche. Nous allons également étudier à travers la traduction la compositionnalité des termes.

Nous abordons dans cette thèse la traduction de langues plus compositionnelles vers des langues qui le sont moins (la composition est plus ou moins productive selon la langue, et en fonction de sa productivité, nous considérons les langues comme plus ou moins compositionnelles). Toutefois, la traduction dans le sens inverse fait également l'objet d'études (Fraser et al., 2012; Stymne et al., 2013).

Ce chapitre commence par un aperçu des méthodes de traduction compositionnelle des unités lexicales, syntagmatiques ainsi que morphologiques (section 7.1). Ensuite nous décrivons nos expériences de traduction des termes identifiés et décomposés par CompoST (section 7.2). Nous effectuons une évaluation quantitative (section 7.3) et qualitative (section 7.4) des résultats obtenus.

7.1 Traduction compositionnelle des termes complexes

L'approche compositionnelle a été abordée dans plusieurs travaux, pour traduire les composés syntagmatiques ainsi que les composés morphologiques et les mots construits morphologiquement. Nous l'avons déjà évoquée dans cette thèse au sujet de l'extraction terminologique multilingue (section 3.2.2). Nous apportons ci-dessous plus de détails.

Cette approche se base sur la possibilité de traduire un composé à partir des traductions de ses composants. Cela présuppose que le sens du composé peut être déduit du sens de ses parties prises individuellement (Baldwin et Tanaka, 2004).

> EN reactive power → FR puissance réactive,
> DE Blindleistungsbedarf 'besoin en puissance réactive' = Blindleistung 'puissance réactive' + Bedarf 'besoin'.

Évidemment tous les composés n'ont pas cette caractéristique. Parfois le sens global ne peut pas être restauré totalement, mais il est toutefois compréhensible dans une certaine mesure (« compositionnalité faible », « *weak compositionality* » selon Pirrelli et al. (2010)) :

> DE Trotzkopf 'personne têtue' = trotzen 'défier' + Kopf 'tête'.

D'autres composés sont complètement opaques ou, autrement dit, idiomatiques. Ainsi, DE *Trotzkopf* a pour équivalent en anglais *pighead* (lit. 'porc' + 'tête'), qui est aussi un composé, mais sémantiquement non transparent.

Il est rarement possible de traduire un composé idiomatique avec l'approche compositionnelle. L'exception est le cas des composés idiomatiques ayant un équivalent dans une autre langue formé avec les mêmes composants sémantiques. Par exemple, l'adjectif russe круглогодичный 'de l'année entière' contient les éléments круглый 'rond' et год 'année', et ne peut pas être traduit en français avec la méthode compositionnelle. Cependant, en anglais il existe des expressions *year-round* et *all-the-year-round* ayant les mêmes composants 'année' et 'rond' qu'en russe, et la traduction serait donc possible.

La méthode compositionnelle a donc ses limites et ne peut pas assurer la traduction de tous les composés. Elle peut toutefois servir pour en traduire une partie. Il a été montré que les composés les plus productifs sont des constructions compositionnelles, ou tout du moins faiblement compositionnelles (Pirrelli et al., 2010). La proportion de composés « compositionnels » dans une langue est difficilement chiffrable, entre autres parce que la compositionnalité n'est pas binaire mais représente plutôt un continuum. Il existe malgré tout certaines estimations. Ainsi, Baldwin et Tanaka (2004) ont calculé sur leurs jeux de tests contenant des composés syntagmatiques nominaux anglais et japonais, pour combien de composés syntagmatiques une traduction produite par un des systèmes de traduction testés correspond à une traduction attestée dans un lexique bilingue contenant des expressions multi-mots. Le taux de ces expressions pour lesquelles on est sûr qu'elles peuvent être traduites compositionnellement dans la langue cible - appelons les « traduisibles compositionnellement » (les auteurs utilisent la dénomination « *translation-compositional data* ») - est de 39 % pour le sens japonais-anglais et de 36 % pour le sens inverse. Robitaille et al. (2006) ont estimé dans leur expérience qu'un ordre de 75 % de composés syntagmatiques français spécialisés peuvent être traduits en japonais compositionnellement.

Ces observations forment « les raisons d'être » des diverses méthodes de traduction compositionnelle. Nous résumons ci-dessous quelques travaux de l'état de l'art dans ce domaine. Dans toutes ces approches on peut distinguer deux étapes principales : la génération des traductions candidates et la sélection des candidates valides.

7.1.1 Traduction des composés syntagmatiques

Dans les travaux sur la traduction compositionnelles, il est d'usage d'employer les termes « unités multi-mots » « *MWU: multi-word units* » ou « expressions multi-mots » « *MWE: multi-word expressions* » plutôt que « composés syntagmatiques ». Pour rester cohérents avec la terminologie employée dans cette thèse, nous allons continuer à utiliser la dénomination « composés syntagmatiques ». Par souci de concision et pour ne pas confondre avec les composés morphologiques, nous introduisons l'acronyme **CS**.

Génération

La génération des traductions candidates d'un CS se fait à l'aide d'un dictionnaire bilingue. Pour illustrer, prenons l'exemple du (Grefenstette, 1999) : nous voulons traduire l'expression française *groupe de travail* vers l'anglais. Dans le dictionnaire FR-EN le mot *groupe* a les traductions suivantes : *cluster, group, concern, grouping, collective*, qui forment un ensemble de cinq éléments (T_1). Le mot FR *travail* est traduit par les mots EN *work, labor, labour*, qui forment un ensemble de trois éléments (T_2). Toutes les combinaisons possibles d'un élément de l'ensemble T_1 avec tout autre élément de l'ensemble T_2 sont ensuite générées, y compris en inversant l'ordre des composants, e.g. *work group*, *labor group*, *grouping work*, etc.

FR groupe → EN cluster, group, concern, grouping, collective,
FR travail → EN work, labor, labour,
FR groupe de travail → EN work cluster, work group, grouping work, ...

Pour limiter le nombre de candidats générés, des patrons structurels prédéfinis peuvent être utilisés (Tanaka et Baldwin, 2003; Baldwin et Tanaka, 2004). Par exemple, pour traduire un CS japonais de structure [N_{s1} + N_{s2}] [1] en anglais, chaque mot de cette expression est traduit individuellement, et seuls les candidats correspondant aux patrons [N_{t1} + in + N_{t2}] ou [N_{t1} + N_{t2}] sont générés.

Un pré-traitement supplémentaire peut consister à découper des CS contenant plus de deux mots graphiques en groupes de mots (Robitaille et al., 2006). Ainsi, l'expression FR *base de données relationnelles* est découpée en : ([base de] et [données relationnelles]), ou ([base de données] et [relationnelles]), ou ([base] et [de données relationnelles]), ou ([base] et [données] et [relationnelles]). Tous ces éléments sont ensuite recherchés dans le dictionnaire bilingue, et certains groupes peuvent avoir plusieurs traductions :

FR base de données → EN database, basis,
FR relationnelles → EN relationnal, related,
FR ([base de données] et [relationnelles]) → EN relational database, relationnal basis, related database, ...

Sélection

Cette étape consiste à choisir les traductions correctes parmi les candidates générées à l'étape précédente. Les candidats sont validés à l'aide du corpus de textes (Tanaka et Baldwin, 2003) ou du web (Grefenstette, 1999). Robitaille et al. (2006) utilisent les corpus de textes collectés sur le web, et ils recherchent les traductions candidates non pas directement dans le corpus cible, mais dans une liste de termes extraits préalablement du corpus.

Si plusieurs candidats sont attestés dans le corpus/web, ils sont généralement classés par fréquence. Ainsi, dans l'exemple ci-dessus tiré du (Grefenstette, 1999) l'expression EN *work group* a une fréquence supérieure à celle des autres candidats, elle est par conséquent considérée comme la meilleure traduction de FR *groupe de travail*.

1. N_{s1} signifie le premier nom de l'expression de la langue source, N_{t1} signifie le premier nom de l'expression de la langue cible.

Des méthodes de classement plus sophistiquées sont proposées. Vintar (2010) assigne à chaque traduction d'un composant du CS source une probabilité estimée à partir d'un corpus parallèle. Le score d'une traduction candidate de ce CS est calculé comme la moyenne arithmétique des probabilités de tous ses composants. Baldwin et Tanaka (2004) utilisent un apprentissage automatique s'appuyant sur des caractéristiques variées issues du corpus cible et du lexique bilingue.

Évaluation de la traduction compositionnelle

Les résultats de la traduction compositionnelle sont souvent évalués par la précision, i.e. le taux de CS traduits correctement parmi tous les CS traduits. Le rappel pour la traduction automatique en général est calculé comme le nombre de traductions correctes proposées par le système, divisé par le nombre d'unités à traduire. Pour la traduction compositionnelle, Robitaille et al. (2006) définissent le rappel comme le nombre de traductions correctes proposées par le système, divisé par le nombre de CS sources pour lesquels une traduction correcte existe dans la liste de CS cibles extraits. Le dénominateur est alors très coûteux à mesurer, surtout avec l'élargissement de la liste de CS cibles, et cela devient quasiment impossible si on utilise non pas la liste de CS extraits mais tout le corpus, voire le web.

Grefenstette (1999) avec sa méthode atteint la précision de 87 % pour la traduction DE-EN et de 86 % pour espagnol-anglais, pour le Top 1. Baldwin et Tanaka (2004) attestent une précision de 84 % dans le sens japonais-anglais et de 78 % pour anglais-japonais pour le Top 1, uniquement sur des CS nominaux et traduisibles compositionnellement. Robitaille et al. (2006) obtiennent une précision d'environ 81 % pour la traduction du français vers le japonais, sur des CS spécifiques à un domaine. Vintar (2010) rapporte une précision moyenne de 83 % pour la traduction slovène-anglais des CS ayant le statut terminologique.

7.1.2 Traduction des mots construits morphologiquement

L'approche compositionnelle a été également utilisée pour traduire des unités complexes contenant un seul mot graphique, c'est-à-dire, les composés morphologiques, les mots préfixés, les dérivés, etc. Le traitement de telles unités nécessite une étape de segmentation préalablement aux deux étapes décrites auparavant. En outre, les composants de ces unités ne se trouvent généralement pas dans les dictionnaires bilingues, des sources lexicales complémentaires sont alors exigées.

Le travail de Cartoni (2009) aborde la traduction des néologismes préfixés de la langue italienne vers le français. Pour identifier les mots préfixés parmi les mots inconnus du dictionnaire, l'auteur construit des règles bilingues de préfixation, e.g. $riV_{it} \rightarrow reV_{fr}$, IT *riconstruire* $\rightarrow$ FR *reconstruire*. Une fois la règle applicable identifiée, il vérifie si la base correspond à une unité lexicale autonome dans un lexique monolingue. Les deux lexiques monolingues s'appuient sur deux larges bases de données morphologiques pour les langues concernées. De plus, ils sont étendus à l'aide de règles reliant des adjectifs relationnels italiens avec leurs bases nominales, et des noms déverbaux avec les verbes. Un lexique bilingue est utilisé pour traduire la base source, en tenant compte des restrictions sur la catégorie grammaticale introduites par la règle de préfixation appliquée. Le web est utilisé pour évaluer la traduction : une traduction candidate est considérée fiable si elle apparaît au moins 5 fois sur le web. La précision de la traduction varie entre 42 % et 94 % selon le préfixe.

Claveau et Kijak (2010) proposent une approche originale à la traduction des termes néoclassiques du français vers le japonais en passant par l'écriture en kanjis. La particularité des caractères kanjis est que chaque signe est porteur du sens, et ils forment des mots par concaténation. En s'appuyant sur ces caractéristiques, les auteurs font l'hypothèse qu'un caractère kanji peut être mis en correspondance avec un morphe français (pour la définition de morphe, cf. section 5.2.4). Pour segmenter les termes français en morphes, ils utilisent donc un ensemble de termes français-japonais alignés extrait d'un thésaurus multilingue, et ils établissent pour chaque terme l'alignement au niveau de morphes. Pour cela, un algorithme d'apprentissage

automatique est exploité, étant entraînés sur un sous-ensemble de termes alignés au niveau de morphes manuellement. La précision de la segmentation obtenue dépasse 70 %. L'algorithme produit aussi une liste de paires « morphe français - kanji » avec la probabilité d'alignement. Ensuite, des nouveaux termes français sont segmentés et traduits en japonais en utilisant cette liste et un parcours de graphes pondérés. La précision de la traduction sur les termes composés est égale à 63 %.

Weller et Heid (2012) effectuent la segmentation des termes composés allemands et leur alignement avec des termes anglais composés ou complexes extraits préalablement d'un corpus anglais. Ils évaluent deux stratégies d'alignement : une première nommée « sac-des-mots » (« *bag-of-words* ») et une seconde basée sur les patrons syntaxiques. La stratégie « sac-des-mots » consiste à générer des traductions candidates avec toutes les permutations possibles des traductions des composants (méthode de base dans la traduction compositionnelle) et à trouver des équivalents des candidates générés dans la liste de termes cibles (les mots grammaticaux étant ignorés en établissant l'équivalence). La deuxième stratégie tient compte de l'ordre des composants pour produire des traductions candidates. En s'appuyant sur l'analyse des traductions obtenues avec la première stratégie, les auteurs sélectionnent sept patrons syntaxiques de traduction, e.g. [N1 N2] → [N2 PREP N1]. Les patrons extraits sont donc adaptés au couple de langues traitées, tandis que la première stratégie ne dépend pas des langues. Avec la première stratégie la précision obtenue est 56 %, et avec la deuxième - 74 % sur le domaine de l'énergie éolienne.

Delpech (2013) traduit des termes construits morphologiquement de l'anglais vers le français et l'allemand. La décomposition des mots en morphèmes est décrite dans la partie 6.3.3. Chaque morphème est traduit dans la langue cible, à savoir que les morphèmes libres sont reliés avec des morphèmes libres, et des morphèmes liés sont reliés soit avec des morphèmes liés, soit avec des morphèmes libres : e.g. EN *post* peut être traduit vers le FR soit comme *post-*, soit comme *après*. Toutes les permutations possibles de traductions des composants sont générées, et ensuite validées selon leur présence dans un corpus de la langue cible. Une méthode de classement est exploitée pour ordonner les traductions validées. L'auteur renforce sa méthode avec l'utilisation de ressources linguistiques variées, notamment des dictionnaires de synonymes et de cognats. Pour la traduction EN→FR des termes relatifs au domaine du cancer du sein, le rappel atteint 62 % et la précision 94 % maximum (les résultats varient en fonction des ressources utilisées). Pour EN→DE, le rappel maximal est de 64 % et la précision de 88 %.

Weller et al. (2014) utilisent l'approche distributionnelle pour estimer automatiquement le degré de compositionnalité d'un composé. L'idée est que les composés avec du sens compositionnel apparaissent dans les textes dans des contextes similaires à ceux des leurs composants, ce qui n'est pas vrai pour des composés idiomatiques. Par exemple, le composé transparent *Holzzaun* 'clôture en bois' apparaît dans des contextes proches à son composant *Zaun* 'clôture' et parfois avec des mêmes co-occurrences que son autre composant *Holz* 'bois', tandis que le composé idiomatique *Jägerzaun* 'clôture en treillis' (lit. 'chasseur + clôture') a des co-occurrences communes avec *Zaun*, mais beaucoup moins avec *Jäger* 'chasseur'. Les valeurs de similarité entre les vecteurs de co-occurrences des composants potentiels et du mot entier sont calculées, et un seuil de similarité est proposé à partir duquel le composé est validé comme compositionnel. Cette approche a été testée dans le cadre de TAS de l'allemand vers l'anglais, mais elle n'a cependant pas abouti à une amélioration significative de la traduction par rapport à l'utilisation de toutes les segmentations générées par une approche équivalente à (Koehn et Knight, 2003) sans distinction entre les composés plus ou moins transparents. Jusqu'à présent, la compositionnalité demeure difficile à capturer avec des méthodes automatiques.

Nous constatons que l'approche compositionnelle de traduction fonctionne plus ou moins bien en fonction des couples de langues et de la méthode utilisée. Malgré des différences dans les résultats obtenus, cette approche s'est montrée utile pour la traduction de composés syntagmatiques ainsi que des mots simples construits morphologiquement et des composés morphologiques. Nous allons appliquer cette approche pour la traduction des termes composés (morphologiques), qui ont été identifiés et décomposés par notre système de segmentation.

7.2 Traduction des termes segmentés par CompoST

La traduction compositionnelle peut servir d'application pour l'évaluation extrinsèque de notre système de segmentation. Pour cela nous allons donner en entré à un module de traduction les segmentations produites par CompoST et évaluer les résultats. Nous avons sélectionné uniquement les termes qui n'ont pas de traduction dans les dictionnaires bilingues utilisés. En outre, cette expérience permettra de donner une approximation de la quantité des termes composés (morphologiques) traduisibles compositionnellement.

Méthode. Nous adoptons une méthode générique qui peut être qualifiée d'approche compositionnelle de base, sans mettre en œuvre les améliorations étroitement liées à un couple de langues spécifique ou aux ressources utilisées. Cela nous permettra de l'appliquer à toutes les langues traitées dans nos expériences.

Génération. Afin de générer les traductions candidates d'un terme composé, le lemme de chaque composant de la segmentation produite par CompoST est d'abord traduit individuellement à l'aide d'un dictionnaire bilingue. Les traductions de tous les composants sont ensuite combinées selon toutes les permutations possibles et concaténées dans une unité soit sans utilisation de séparateur, soit avec un espace ou un tiret entre les éléments. Le module de segmentation peut proposer plusieurs segmentations pour un mot. Dans ce cas, il est possible d'utiliser soit la segmentation la mieux classée, soit générer tous les candidats à partir des N meilleures segmentations. Nous nous limitons ici à utiliser la segmentation la mieux classée (Top 1 de segmentation).

Sélection. Nous introduisons ici une seule modification par rapport à l'approche de base au niveau de la sélection des candidats. Pour valider les traductions candidates, nous les cherchons non dans le corpus cible directement, mais dans une liste de termes extraits du corpus en amont à l'aide de l'extracteur TermSuite. L'utilisation de cette liste de termes permet de filtrer les candidats polylexicaux erronés qui ne forment pas un syntagme.

Ce procédé nous permet d'aligner un composé morphologique de la langue source soit avec un composé syntagmatique, soit avec un composé morphologique de la langue cible :

DE Aromatasehemmer = aromatase + hemmer → EN aromatase inhibitor;
EN chemosensitivity = chemo + sensitivity → FR chimiosensibilité.

La méthode de traduction est schématisée sur la figure 7.1.

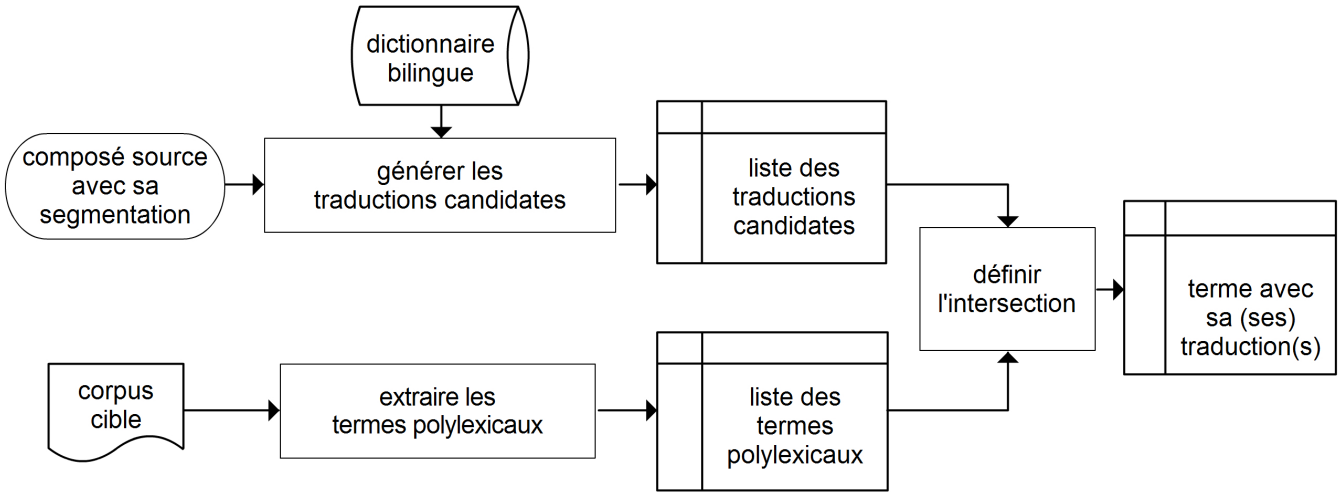

FIGURE 7.1 – Méthode de traduction compositionnelle

Données. Nous réalisons la traduction selon cette approche pour les paires de langues suivantes : RU → EN (domaine de l'énergie éolienne), EN → FR (pour les deux domaines de spécialité), DE → EN (pour les deux domaines de spécialité). Nous réutilisons des ressources présentées dans l'annexe B : (1) les NCP-listes bilingues pour traduire des éléments néoclassiques des termes composés, (2) les dictionnaires bilingues pour traduire des composants natifs et (3) les corpus spécialisés pour extraire des termes complexes. Nous rappelons que les corpus utilisés sont comparables ce qui devrait assurer la présence des termes équivalents dans les corpus de langues différentes (Morin et al., 2007).

Expériences. Nous appliquons CompoST à la liste des noms et des adjectifs (étiquetés et lemmatisés par TreeTagger) issus d'un corpus spécialisé monolingue. Nous faisons l'hypothèse que la segmentation effectuée avec le seuil optimisé pour le rappel sera plus utile pour la tâche de traduction compositionnelle que celle faite avec le seuil optimisé pour la précision, car les fausses segmentations candidates seront filtrées par le module de traduction. Les systèmes avec une tendance à la sur-segmentation se sont déjà montrés plus efficaces par rapport à ceux avec une tendance à la sous-segmentation dans la traduction automatique statistique à base de segments (« *phrase-based statistical machine translation* ») (Koehn et Knight, 2003; Weller et al., 2014). Pour vérifier si cette hypothèse est aussi vraie pour la traduction compositionnelle, nous effectuons la traduction avec les deux options (optimisant le rappel et la précision) et nous les comparons.

Afin d'estimer la limite supérieure de performance que nous pourrons atteindre pour une langue donnée et avec les ressources à notre disposition, nous effectuons aussi la traduction en exploitant les segmentations de référence utilisées pour l'évaluation intrinsèque de la segmentation, et nous comparons les résultats à ceux obtenus dans les expériences décrites ci-dessus.

Dans toutes ces expériences, les traductions obtenues sont évaluées manuellement.

7.3 Évaluation quantitative

7.3.1 Stratégies d'évaluation

Nous évaluons la qualité de la traduction obtenue avec les mesures habituelles : la précision, le rappel et la F-mesure. Pour les calculer, nous envisageons deux stratégies possibles :

1. **Évaluation sur des termes composés.** La traduction compositionnelle est généralement évaluée sur l'ensemble des unités notoirement composées ou construites morphologiquement. Pour pouvoir comparer nos résultats à d'autres travaux, nous calculons la précision et le rappel sur les termes composés seulement (en écartant les termes non-composés même s'ils ont été segmentés et traduits). Pour cela, la précision P_{comp} est calculée comme la taille de l'intersection entre les composés traduits C_{TR} et les traductions de référence des termes composés C_R, divisée par le nombre de composés traduits :

$$P_{comp} = \frac{|C_{TR} \cap C_R|}{|C_{TR}|} \tag{7.1}$$

Le rappel R_{comp} est calculé comme la taille de l'intersection entre les composés traduits C_{TR} et les traductions de référence C_R, divisée par le nombre de traductions de référence :

$$R_{comp} = \frac{|C_{TR} \cap C_R|}{|C_R|} \tag{7.2}$$

La F-mesure est calculée comme la moyenne harmonique entre la précision et le rappel.

2. **Évaluation sur des termes segmentés.** Dans une application réelle, si la segmentation est utilisée dans la traduction de manière non supervisée par l'utilisateur, toutes les unités segmentées seront potentiellement traduites, y compris celles segmentées par erreur. Pour reproduire cette situation, nous évaluons la traduction faite en utilisant toutes les segmentations produites par CompoST, y compris les fausses positives. Dans ce cas, la précision $P_{comp+non}$ est calculée comme la taille de l'intersection entre les termes traduits T_{TR} (les composés et les non-composés) et les traductions de référence T_R, divisée par le nombre de termes traduits :

$$P_{comp+non} = \frac{|T_{TR} \cap T_R|}{|T_{TR}|} \tag{7.3}$$

Le rappel $R_{comp+non}$ est calculé comme la taille de l'intersection entre les termes traduits T_{TR} (les composés et les non-composés) et les traductions de référence T_R, divisée par le nombre de termes segmentés T_S :

$$R_{comp+non} = \frac{|T_{TR} \cap T_R|}{|T_S|} \tag{7.4}$$

Les traductions de référence dans ce cas comprennent les traductions de tous les termes segmentés, par conséquent $|T_R| = |T_S|$.

Nous appliquons les deux stratégies pour évaluer nos résultats.

7.3.2 Évaluation de la traduction des termes composés

Avec la segmentation optimisée pour le rappel. Le tableau 7.1 présente les résultats de la traduction compositionnelle pour les termes composés, la segmentation étant faite avec le seuil optimisé pour le rappel.

TABLE 7.1 – Évaluation de la qualité de la traduction compositionnelle des termes composés (segmentés par CompoST avec le seuil optimisé pour le rappel).
P signifie la précision, R - le rappel, F - F-mesure, C_{TR} - les composés traduits, C_R - les traductions de référence des termes composés.

	ÉNERGIE ÉOLIENNE			CANCER DU SEIN	
	RU-EN	EN-FR	DE-EN	EN-FR	DE-EN
$\|C_R\|$	170	100	103	201	109
$\|C_{TR}\|$	41	30	42	50	35
$\|C_{TR} \cap C_R\|$	34	19	28	46	31
P_{comp} (%)	83	63	67	92	89
R_{comp} (%)	20	19	27	23	28
F_{comp} (%)	32	29	39	37	43

La précision sur l'ensemble des termes composés varie selon la langue et le domaine entre 63 % (EN-FR énergie éolienne) et 92 % (EN-FR cancer du sein). Le rappel est compris entre 19 % (EN-FR énergie éolienne) et 28 % (DE-EN cancer du sein). La F-mesure est comprise entre 29 % (EN-FR énergie éolienne) et 43 % (DE-EN cancer du sein).

Résultats par domaine. Les résultats les plus faibles sont obtenus pour EN-FR dans le domaine de l'énergie éolienne. Une des raisons à cela est le taux élevé de quasi-composés en anglais qui sont particulièrement difficiles à traduire compositionnellement. Une autre raison est le grand nombre d'adjectifs dans ce jeu de

données (e.g. EN *cost-effective, cross-sectional* etc.) qui ne se traduisent pas en français de manière compositionnelle : *bon rapport coût/efficacité* (traduction fertile), *transversal* (non segmentable en composants en synchronie). Remarquons que pour le même couple de langues, mais sur le domaine du cancer du sein, nous obtenons la précision la plus élevée, ce qui confirme que l'impact du domaine est très important.

Résultats par type de phénomène. Nous avons également analysé les résultats de traduction pour chaque catégorie de termes traités séparément (cf. notre typologie opérationnelle, section 5.1.5). Les résultats sont résumés dans le tableau 7.2. La précision et le rappel ont été calculés uniquement sur les composés (avec le seuil optimisé pour le rappel), étant donné que dans cette évaluation nous nous sommes intéressée aux proportions de termes de chaque catégorie traduits correctement.

Les composés néoclassiques sont traduits correctement si leurs éléments sont présents dans nos NCP-listes. En effet, la précision sur les NC pris séparément est égale à 100 % pour toutes les expériences, tandis que pour les natifs la précision n'a pas dépassé 85 %. Les mots préfixés sont un peu moins facilement traduisibles et la présence du préfixe dans la liste bilingue est une condition primordiale de traduction. Les quasi-composés ne posent pas de problème pour la segmentation, mais ils sont difficiles à traduire compositionnellement et hors contexte : par exemple EN *case-control* généralement utilisé dans l'expression *case-control study* 'étude cas-témoins', a été traduit en FR comme *cas du contrôle*. Les emprunts sont trop peu nombreux dans nos données pour faire des observations quantitatives sur leurs traductions.

Avec la segmentation optimisée pour la précision. Notre hypothèse était que la segmentation effectuée avec les paramètres optimisés pour le rappel serait plus utile pour la tâche de traduction compositionnelle que celle optimisée pour la précision. Pour vérifier cette hypothèse, nous répétons les expériences de traduction en utilisant les segmentations produites par CompoST optimisé pour la précision. Les résultats (sur les termes composés) sont présentés dans le tableau 7.3 et devront être comparés avec le tableau 7.1.

Avec la segmentation optimisée pour la précision, nous obtenons moins de segmentations, et par conséquent, moins de traductions candidates. Cependant, le nombre de traductions correctes parmi ces candidates est aussi inférieur. Finalement, pour EN-FR et DE-EN (domaine de l'énergie éolienne) la F-mesure de la traduction reste égale à celle obtenue avec la segmentation optimisée pour le rappel, pour les mêmes couples de langues et le domaine médical la F-mesure est inférieure (l'écart de 2 points), et pour RU-EN (énergie éolienne) l'écart est assez important (14 points) en faveur de la traduction faite avec la segmentation optimisée pour le rappel. Nous constatons donc que notre hypothèse se confirme sur nos données, au moins dans l'évaluation faite sur les termes composés. Nous allons la vérifier également dans l'évaluation sur tous les termes segmentés.

Avec les segmentations de référence. Nous comparons les résultats obtenus en utilisant CompoST paramétré pour optimiser le rappel (cf. tableau 7.1) avec ceux obtenus en utilisant les segmentations de référence exploitées dans l'évaluation intrinsèque de la segmentation (cf. tableau 7.4). Cela permettra de démontrer les limites les plus hautes que nous pourrions obtenir avec la méthode de traduction et les ressources lexicales utilisées.

La F-mesure obtenue pour la traduction effectuée avec les segmentations de référence varie entre 29 % (EN-FR énergie éolienne) et 45 % (DE-EN cancer du sein). Nous constatons que les résultats obtenus avec les segmentations produites par CompoST les suivent de près : pas d'écart pour le couple EN-FR (sur les deux domaines), 2 points d'écart pour DE-EN (cancer du sein), 3 points pour les deux autres expériences. Ce résultat est plutôt encourageant pour notre méthode de segmentation. Le fait que les résultats ne sont pas élevés même avec l'utilisation des segmentations correctes uniquement s'explique, d'un côté, par la non-compositionnalité de certains composés, et de l'autre côté, par l'incomplétude des ressources lexicales (cf. l'évaluation qualitative). Le rappel attesté avec l'utilisation des segmentations de référence montre combien de termes de la langue source, parmi ceux sélectionnés, sont traduisibles suite à la segmentation.

TABLE 7.2 – Évaluation de la qualité de la traduction compositionnelle en fonction du type de composé. P signifie la précision, R - le rappel, NV signifie composés natifs, NC - composés néoclassiques, PR - mots préfixés, QC - quasi-composés, LN - composés empruntés.

	NV	NC	PR	QC	LN
RU-EN ÉNERGIE ÉOLIENNE					
Nb composés	98	53	16	0	3
Nb traductions	28	9	5	0	0
Nb traductions correctes	23	9	3	0	0
P (%)	82	100	60	0	0
R (%)	24	17	19	0	0
EN-FR ÉNERGIE ÉOLIENNE					
Nb composés	55	14	19	12	0
Nb traductions	23	2	2	2	0
Nb traductions correctes	14	2	2	1	0
P (%)	58	100	100	50	0
R (%)	25	14	11	8	0
EN-FR CANCER DU SEIN					
Nb composés	40	94	40	27	0
Nb traductions	11	29	9	1	0
Nb traductions correctes	9	29	8	1	0
P (%)	82	100	89	100	0
R (%)	23	30	20	4	0
DE-EN ÉNERGIE ÉOLIENNE					
Nb composés	89	2	4	0	8
Nb traductions	37	1	1	0	3
Nb traductions correctes	26	1	0	0	1
P (%)	70	100	0	0	33
R (%)	29	50	0	0	12
DE-EN CANCER DU SEIN					
Nb composés	73	32	3	0	1
Nb traductions	27	8	0	0	0
Nb traductions correctes	23	8	0	0	0
P (%)	85	100	0	0	0
R (%)	32	25	0	0	0

TABLE 7.3 – Évaluation de la qualité de la traduction compositionnelle des termes composés (segmentés par CompoST avec le seuil optimisé pour la précision).
P signifie la précision, R - le rappel, F - F-mesure, C_{TR} - les composés traduits, C_R - les traductions de référence des termes composés.

	ÉNERGIE ÉOLIENNE			CANCER DU SEIN	
	RU-EN	EN-FR	DE-EN	EN-FR	DE-EN
$\|C_R\|$	170	100	103	201	109
$\|C_{TR}\|$	23	30	39	46	33
$\|C_{TR} \cap C_R\|$	17	19	28	43	29
P_{comp} (%)	74	63	72	93	88
R_{comp} (%)	10	19	27	21	27
F_{comp} (%)	18	29	39	35	41

TABLE 7.4 – Traduction compositionnelle des termes composés avec la segmentation de référence.
P signifie la précision, R - le rappel, F - F-mesure, C_{TR} - les composés traduits, C_R - les traductions de référence des termes composés.

	ÉNERGIE ÉOLIENNE			CANCER DU SEIN	
	RU-EN	EN-FR	DE-EN	EN-FR	DE-EN
$\|C_R\|$	170	100	103	201	109
$\|C_{TR}\|$	43	29	44	50	38
$\|C_{TR} \cap C_R\|$	37	19	31	47	33
P_{comp} (%)	86	66	70	94	87
R_{comp} (%)	22	19	30	23	30
F_{comp} (%)	35	29	42	37	45

Nous pouvons interpréter cette valeur comme étant la limite basse de compositionnalité dans une langue (au moins cette proportion de termes peut être considérée comme compositionnelle). Il est cependant plus judicieux de l'interpréter comme une valeur de *traduisibilité compositionnelle* entre deux langues, car la possibilité de traduire un terme dépend de la compositionnalité de son équivalent dans la langue cible tout autant que de sa compositionnalité dans la langue source. Ainsi, pour les termes RU-EN cette proportion s'avère égale à 22 % pour le domaine de l'énergie éolienne, pour EN-FR - à 19 % pour l'énergie éolienne et à 23 % pour le cancer du sein, et pour DE-EN à 30 % pour les deux domaines.

Comparaison avec des résultats de l'état de l'art. Il est difficile de comparer les résultats obtenus ici avec ceux rapportés dans les travaux sur la traduction compositionnelle du fait que : (1) des travaux différents ne délimitent pas de la même manière les phénomènes à traiter : des composés syntagmatiques, des mots préfixés, des composés néoclassiques, etc., (2) les ressources et les paires de langues utilisées dans les expériences sont différentes.

Parmi les travaux cités dans la partie 7.1 d'état de l'art, ceux de (Weller et Heid, 2012) et (Delpech, 2013) sont les plus proches de nos expériences en ce qui concerne les paires de langues ainsi que les phénomènes traités. Les deux effectuent la segmentation et ensuite la traduction compositionnelle de l'ensemble des termes composés (ainsi que des termes formés par la suffixation et la préfixation dans le cas de (Delpech, 2013)). Par conséquent, leurs résultats doivent être comparés avec notre évaluation effectuée sur les termes composés (le tableau 7.1).

Weller et Heid (2012) attestent une précision d'environ 56 % pour l'alignement sans patrons syntaxiques et de 74 % pour l'alignement à l'aide des patrons pour DE→EN sur le domaine de l'énergie éolienne. Pour

l'expérience analogue, nous obtenons une précision de 67 % sans utilisation de patrons syntaxiques.

Pour EN→FR sur le corpus du cancer du sein Delpech (2013) atteint la précision de 94 % et le rappel de 62 %. Nous obtenons pour cette paire de langues et ce domaine une précision proche (92 %) et un rappel nettement plus faible (23 %). Une telle différence de rappel s'explique en priorité par la différence d'approches. Le travail de Delpech (2013) est focalisé sur la traduction, la segmentation étant seulement la première étape du traitement. L'auteur choisit les termes composés de manière à ce que tous leurs composants non indépendants (préfixes, suffixes, racines néoclassiques) soient présents dans la liste bilingue utilisée. De plus, l'auteur utilise plusieurs ressources linguistiques complémentaires y compris le dictionnaire des cognats, très utile pour la traduction dans les domaines médicaux. En utilisant uniquement un dictionnaire bilingue et une liste bilingue d'éléments non indépendants, elle obtient la précision de 93 % et le rappel de 34 %, ce qui est plus proche des résultats produits par CompoST.

Nous voulons également rappeler que nous avons utilisé pour la traduction une méthode assez basique. Notre but n'était pas d'améliorer les méthodes existantes de traduction compositionnelle, mais d'évaluer à quel point la segmentation obtenue avec CompoST est utile pour cette tâche, pour les paires de langues et les domaines traités.

7.3.3 Évaluation de la traduction des termes segmentés

Nous évaluons maintenant la traduction des tous les termes segmentés, comprenant les termes composés et non-composés.

Les résultats obtenus avec **la segmentation optimisée pour le rappel** sont présentés dans le tableau 7.5. La précision varie selon la langue et le domaine entre 64 % (EN-FR énergie éolienne) et 89 % (EN-FR énergie éolienne, DE-EN cancer du sein), le rappel varie entre 14 % (RU-EN, EN-FR énergie éolienne) et 28 % (DE-EN cancer du sein) et la F-mesure entre 23% (RU-EN, EN-FR énergie éolienne) et 42 % (DE-EN cancer du sein). Par rapport à l'évaluation uniquement sur les termes composés (tableau 7.1), nous constatons ici les mêmes régularités, sauf que les valeurs de toutes les mesures sont généralement moins élevées. La différence est moins perceptible pour la traduction DE-EN. Nous nous rappelons que la segmentation en allemand générait largement moins de faux positifs que dans les autres langues. De ce fait, très peu de fausses segmentations DE ont abouti à une traduction.

Les résultats obtenus avec **la segmentation optimisée pour la précision** sont présentés dans le tableau 7.6. Nous constatons qu'avec la segmentation optimisée pour la précision dans chaque expérience le nombre de termes traduits est inférieur au nombre de termes traduits avec la segmentation optimisée pour le rappel, de même pour le nombre de traductions correctes. Le nombre de termes segmentés est aussi inférieur, avec un écart plus ou moins grand (de 4 unités pour DE cancer du sein jusqu'à 163 unités pour RU).

TABLE 7.5 – Évaluation de la qualité de la traduction compositionnelle des termes segmentés par CompoST (avec le seuil optimisé pour le rappel).
P signifie la précision, R - le rappel, F - F-mesure, T_{TR} - les termes traduits, T_R - les traductions de référence, T_S - les termes segmentés.

	ÉNERGIE ÉOLIENNE			CANCER DU SEIN	
	RU-EN	EN-FR	DE-EN	EN-FR	DE-EN
$\|T_S\|$	271	154	120	281	112
$\|T_{TR}\|$	50	33	46	57	35
$\|T_{TR} \cap T_R\|$	37	21	30	51	31
$P_{comp+non}$ (%)	74	64	65	89	89
$R_{comp+non}$ (%)	14	14	25	18	28
$F_{comp+non}$ (%)	23	23	36	30	42

TABLE 7.6 – Évaluation de la qualité de la traduction compositionnelle des termes segmentés par CompoST (avec le seuil optimisé pour la précision).
P signifie la précision, R - le rappel, F - F-mesure, T_{TR} - les termes traduits, T_R - les traductions de référence, T_S - les termes segmentés.

	ÉNERGIE ÉOLIENNE			CANCER DU SEIN	
	RU-EN	EN-FR	DE-EN	EN-FR	DE-EN
$\|T_S\|$	108	117	105	202	98
$\|T_{TR}\|$	28	33	41	48	33
$\|T_{TR} \cap T_R\|$	20	21	29	45	29
$P_{comp+non}$ (%)	71	64	71	94	88
$R_{comp+non}$ (%)	19	18	28	22	30
$F_{comp+non}$ (%)	29	28	40	36	44

Finalement la F-mesure s'avère plus haute avec cette option pour tous les couples de langues. Cela veut dire que dans le contexte d'utilisation non-supervisée de la segmentation automatique pour la traduction compositionnelle, notre hypothèse, selon laquelle la segmentation optimisée pour le rappel est plus performante, ne se confirme pas, et la segmentation optimisée pour la précision permet d'obtenir la meilleure traduction en termes de F-mesure.

7.4 Évaluation qualitative

Nous analysons les erreurs fréquentes de traduction compositionnelle sur l'exemple de la traduction de tous les termes segmentés par CompoST paramétré pour optimiser le rappel (le nombre d'erreurs produites suite à la segmentation optimisée pour la précision est moindre). Les cas d'échec de traduction peuvent être divisés en deux types : soit un composé n'a pas été traduit, soit un mot a été traduit incorrectement.

7.4.1 Absence de traduction

Les causes principales d'absence de traduction des composés sont les suivantes :

1. Sous-segmentation. Nous savons que tous les termes composés n'ont pas été segmentés (cf. le rappel de segmentation). Un terme qui n'a pas été segmenté ne peut pas être traduit compositionnellement.
2. Insuffisance des ressources lexicales. Même si le composé est bien segmenté, il ne peut pas être traduit si un de ses composant manque dans le dictionnaire ou NCP-liste.
3. Insuffisance du corpus cible. Parfois une traduction correcte d'un terme source est générée, mais n'est pas validée parce qu'elle n'apparaît pas dans le corpus cible.
4. Non-compositionnalité. Nos expériences confirment une évidence : la méthode compositionnelle n'est pas adaptée à la traduction des composés dont le sens ne découle pas du sens de ses éléments. Par exemple, EN *soft-starter* ne peut pas être aligné avec FR *démarreur progressif*, parce que le mot *soft* ne se traduit pas par *progressif* dans d'autres contextes.

La première cause est liée à la segmentation. La deuxième et la troisième causes sont relatives à la couverture des ressources. Le dernier cas montre les limites de la méthode compositionnelle.

TABLE 7.7 – Traductions incorrectes et leurs causes (en gris - les erreurs liées à la segmentation).

Langue et domaine	Nb. erreurs total	Sur-segmentation	Segmentation incorrecte	Sur-génération
DE-EN Énergie éolienne	16	2 (13 %)	1 (6 %)	13 (81 %)
EN-FR Cancer du sein	5	2 (40 %)	0	3 (60 %)
RU-EN Énergie éolienne	13	6 (46 %)	2 (15 %)	5 (39 %)

7.4.2 Traduction incorrecte

Les causes de traduction incorrecte peuvent aussi être divisés en trois types :

1. Sur-segmentation. Parfois les mots non-composés sont segmentés par notre méthode de segmentation et ensuite traduits, e.g. EN *operability* a été segmenté en *open* + *ability* et aligné avec l'expression française *pouvoir public* (*open* → *publique, ability* → *pouvoir*).
2. Segmentation incorrecte. Des composés peuvent être segmentés incorrectement, ou un lemme erroné peut être assigné à un des composants. Cela provoque également un alignement impropre. Par exemple, l'adjectif RU длинноволновый 'à grandes ondes', lit. 'longue + onde', a été segmenté en длина 'longueur' + волна 'onde' ce qui est quasiment correct si on omet le changement de catégorie grammaticale (*long* → *longueur*). Ensuite le terme a été aligné avec EN *wavelength* 'longueur d'onde'.
3. Sur-génération (« *over-generation* »). Ceci est un problème typique pour l'approche compositionnelle en traduction. Même les segmentations correctes peuvent engendrer des traductions erronées à cause de polysémie et d'homonymie caractéristiques pour toutes les langues. Par exemple RU многоразовый 'réutilisable' a été segmenté proprement en много 'beaucoup' + раз 'fois'. Le dernier composant étant aligné avec EN *time*, cela a abouti à l'alignement *considerable time*. Cette erreur se produit également à cause de la permutation des composants utilisée pendant la traduction. Ainsi il est connu que dans un composé DE de structure [N1N2] l'élément N2 sert de tête de groupe nominal, et N1 de modificateur, donc le composé doit être traduit en EN comme [N2 of N1] ou [N1 N2], e.g. DE *Kabelsystem* = *Kabel* + *System* → EN *cable system*. Cependant il a été traduit dans notre expérience par *system cable* où *cable* est la tête et *system* est le modificateur, ce qui change complètement le sens du terme.

Le premier et le deuxième types d'erreur sont provoqués par des défauts de segmentation, le dernier est lié à la méthode de traduction utilisée. Le tableau 7.7 montre le taux de chaque type d'erreurs sur les paires langues-domaine suivantes : EN-FR dans le domaine du cancer du sein, RU-EN et DE-EN dans le domaine d'énergie éolienne. Les erreurs liées à la segmentation (la segmentation incorrecte ainsi que la sur-segmentation) représentent de 19 % à 61 % de toutes les erreurs de traduction sur ces expériences.

Le tableau 7.8 montre que parmi ces segmentations incorrectes seul un petit pourcentage a abouti à l'alignement, la majorité des candidats erronés n'ont pas été validés par le corpus cible. Cela a déjà été démontré par Weller et Heid (2012) dans une évaluation à une petite échelle, et s'est confirmé sur nos données. Dans ce tableau on peut aussi voir que même certaines segmentations incorrectes ont été traduites correctement. Cela provient des mots étrangers apparaissant dans nos corpus : par exemple dans le corpus EN du cancer du sein on retrouve le mot FR *mastectomie*, qui a été extrait comme un terme anglais, ensuite segmenté en *mast* + *ectomy* (*ectomie* étant relié au mot EN *ectomy* pendant la segmentation) et aligné avec FR *mastectomie*. Nous considérons cette segmentation en FR comme incorrecte, mais l'alignement qui en résulte comme correct.

TABLE 7.8 – Impact des erreurs de segmentation dans la traduction.

Langue et domaine	Nb. segmentations incorrectes	Traduites correctement	Traduites incorrectement	Filtrées par alignement
DE-EN Énergie éolienne	39	2 (5%)	3 (8%)	34 (87%)
EN-FR Cancer du sein	112	5 (4%)	2 (2%)	105 (93%)
RU-EN Énergie éolienne	143	2 (1%)	8 (6%)	133 (93%)

7.5 Bilan

Nous avons exploité les résultats de la segmentation des termes composés dans la traduction compositionnelle. Nous constatons que cela permet de traduire une partie de termes qui n'ont pas de traduction dans les dictionnaires de langue générale, et de produire des traductions ayant une structure variée (soit terme composé, soit terme complexe) dans la langue cible.

Nous avons évalué les résultats de la traduction selon deux stratégies : uniquement sur les termes composés (afin de comparer aux résultats de l'état de l'art et aux résultats obtenus avec l'utilisation des segmentations de référence) et sur tous les termes segmentés.

Étant donnée la tâche visée, nous avons opté pour la segmentation optimisant le rappel, en supposant que les segmentations non-plausibles seraient filtrées par le module de traduction. En effet, une large partie des segmentations incorrectes ont été filtrées (de l'ordre de 90 % en utilisant la meilleure segmentation), même si nous n'avons pas complètement échappé aux erreurs de traduction liées à la segmentation. Le taux d'erreurs dues à la segmentation varie considérablement en fonction des langues et du domaine (entre 19 % et 61 % de toutes les erreurs, pour trois expériences). Nous avons comparé la traduction effectuée en utilisant la segmentation optimisée pour le rappel avec la traduction effectuée en utilisant la segmentation optimisée pour la précision selon les deux stratégies d'évaluation. Dans l'évaluation portant uniquement sur les termes composés la segmentation optimisant le rappel s'est montrée plus efficace, mais dans l'évaluation sur tous les termes segmentés, qui correspond à un contexte plus réaliste d'utilisation, la segmentation optimisant la précision s'est avérée meilleure en termes de F-mesure. Cela veut dire que la précision relativement élevée de la segmentation automatique est conséquente pour la traduction compositionnelle, contrairement à la TAS.

Sur les termes composés, avec la segmentation optimisée pour le rappel nous avons obtenu une précision entre 63 % et 92 %, et un rappel entre 19 % et 28 % selon la paire de langues et le domaine. Le rappel de traduction ne dépend pas que de la qualité de la segmentation. Tous les composés n'ont pas de sens compositionnel, par conséquent un rappel élevé ne peut pas être obtenu avec l'approche compositionnelle. De plus, parfois le silence s'explique par l'insuffisance des ressources lexicales utilisées et du corpus cible. Nous avons observé que l'écart entre le rappel de la traduction obtenue en utilisant les segmentations faites par CompoST et celle obtenue en utilisant les segmentations de références est relativement petit.

Sur tous les termes segmentés, avec la segmentation optimisée pour le rappel la précision varie entre 64 % et 89 %, et le rappel entre 14 % et 28 %. Cette précision montre à quel point les traductions produites par la méthode compositionnelle de base et en utilisant la sortie de notre méthode de segmentation sont bruitées. Nous constatons que la précision pour le domaine du cancer du sein est relativement élevée (89 % pour EN-FR et pour DE-EN), tandis que la précision pour le domaine de l'énergie éolienne est nettement plus inférieure (de 64 % à 74 %).

Nous nous posons la question sur la possibilité d'utiliser les traductions produites avec ce niveau de bruit. Ici encore tout dépend de l'application finale. Le module de traduction compositionnelle peut faire partie d'un système de traduction assistée par ordinateur. Dans de tels systèmes, les différentes traductions pour un mot ou une expression sont suggérées à l'utilisateur, et c'est à ce dernier d'en choisir une. Dans ce contexte la précision obtenue dans nos expériences est acceptable. La traduction compositionnelle peut être

également utilisée pour améliorer la traduction automatique. Les systèmes TAS traduisent non seulement des mots, mais des phrases entières. Cependant si au moins un mot n'a pas été traduit, la traduction de toute la phrase échoue (Cartoni, 2009). Les mots qui n'ont pas été traduits par un système TAS peuvent être transférés en entrée au module de traduction compositionnelle, puis les traductions produites par ce module sont retournées au système TAS, et la traduction de la phrase est réitérée. Dans ce cadre une haute précision de la traduction compositionnelle est très importante, puisque l'absence de traduction d'un mot laisse la phrase « sous-traduite » ou non traduite, tandis que la traduction incorrecte peut générer la fausse traduction de toute la phrase sans que l'utilisateur le sache (dans l'hypothèse qu'il ne maîtrise pas la langue cible). Nous jugeons donc que la précision obtenue dans nos expériences sur le domaine de l'énergie éolienne n'est pas suffisante pour utiliser les résultats directement dans un système TAS.

Pour augmenter la précision, il faudrait améliorer la technique de traduction compositionnelle. Par exemple, on pourrait introduire des patrons de traduction comme le font Weller et Heid (2012), ou attribuer à chaque traduction d'un composant trouvée dans le lexique une probabilité estimée à l'aide d'un corpus parallèle, comme le fait Vintar (2010). Nous précisons aussi que dans ce travail nous n'avons pas ordonné les traductions candidates produites. Pour pouvoir utiliser les résultats de la traduction compositionnelle dans un système TAS, les candidates doivent être classées par pertinence.

8

Détection des variantes syntagmatiques des termes composés

Comme tous les termes, les termes composés ont des variantes. Le regroupement des variantes d'un terme est utile dans plusieurs applications du TAL telles que la traduction automatique statistique et assistée par ordinateur (Weller et al., 2011), la recherche d'informations (Yoshikane et al., 1999), le remplissage de bases de données terminologiques et la construction de systèmes question-réponse. La variation syntaxique a déjà été étudiée dans de nombreux travaux et des outils pour la capturer ont été mis en place, mais peu de travaux ont traité les variantes syntagmatiques des composés morphologiques. Ce chapitre est consacré à cette problématique.

Dans ce chapitre, nous allons extraire à partir des corpus spécialisés des variantes de type « terme monolexical - terme polylexical » ou autrement dit « composé morphologique (CM) - composé syntagmatique (CS) ». L'objectif est d'étudier les structures des variantes détectées dans une perspective multilingue. De plus, cette expérience fournit un deuxième exemple de l'application de la segmentation automatique pour la construction d'un lexique terminologique, car de tels lexiques pourraient être enrichis par des variantes des termes. Pour identifier les termes composés dans les corpus, nous utilisons CompoST.

Il s'agit ici de la variation dénominative et non conceptuelle (cf. section 2.6). Nous adoptons la définition de (Daille et al., 1996, p. 201) : « la variante d'un terme est un énoncé qui est sémantiquement et conceptuellement lié au terme d'origine ». Nous considérons comme variantes les paires de termes qui ont entre eux une relation de synonymie :

> DE Brustdrüsenentfernung, lit. 'glande + mammaire + ablation' ↔
> Entfernung der Brustdrüse, 'ablation de la glande mammaire',

ou une relation de quasi-synonymie (c'est-à-dire des termes interchangeables mais seulement dans certains contextes), par exemple :

> RU энергоисточник, lit. 'énergie + source' ↔
> источник тепловой энергии, 'source de l'énergie thermique'.

Après avoir résumé quelques travaux concernant l'extraction des variantes et surtout de celles incluant des composés (section 8.1), nous allons mettre en place un algorithme simple et indépendant de la langue afin d'extraire les variantes (section 8.2). Nous décrivons nos expériences et les paramètres des outils impliqués dans la section 8.3 et les résultats obtenus dans la section 8.4. Nous concluons avec la section 8.5.

8.1 Extraction des variantes de termes

La détection des variantes terminologiques dans un contexte multilingue a été examinée pour l'anglais (Jacquemin, 1994; Ville-Ometz et al., 2007), le français (Jacquemin, 1994; Weller et al., 2011), l'allemand (Weller et al., 2011) et le japonais (Yoshikane et al., 1999).

Jacquemin (1994) a conçu un outil de reconnaissance des variantes dans les corpus de textes, Fastr, basé sur des règles morphosyntaxiques. Il s'agit ici des variantes de type CM-CS, par exemple EN *ambiguity resolution* 'résolution d'ambiguïté' : *ambiguities are resolved* 'les ambiguïtés sont résolues', *ambiguity types resolved* 'les types d'ambiguïté sont résolus', *resolution of lexical ambiguity* 'l'ambiguïté lexicale est résolue', etc. (Jacquemin, 1999). Ville-Ometz et al. (2007) ont élargi ces règles pour l'anglais en définissant le contexte gauche et droit d'un groupe nominal qui forme une variante. Pour les variantes de ce type, la précision moyenne s'est avérée égale à 83 %.

Yoshikane et al. (1999) ont adapté Fastr pour la détection des variantes en japonais. Pour établir les règles morphosyntaxiques exigées par Fastr, les auteurs ont aligné des composés syntagmatiques issus d'une base de données terminologique avec des phrases extraites du corpus de textes qui contenaient tous les mots pleins des CS correspondants. L'examen des paires obtenues leur a permis de définir des patrons, ou autrement dit des règles de variation en japonais. Un large spectre de types de variation a été considéré dans ce travail, y compris le type CS-CM pour lequel une haute précision de 94 % a été attestée sur les 806 variantes extraites.

Weller et al. (2011) évoquent des variantes allemandes de type CM-CS dans lesquelles un des composants du CS est étendu avec un élément supplémentaire, formant à son tour un composé morphologique :

DE Nutzenergie 'énergie utile' ↔ nutzbar Energie**form** 'forme d'énergie utilisable',

Cependant dans les expériences rapportées, les auteurs ne traitent qu'un patron de variation pour l'allemand, « [N PREP N] - CM » ne couvrant pas la variation avec expansion présentée dans l'exemple. La précision obtenue est de 74 % sur 100 variantes allemandes. Les auteurs proposent également une autre approche d'extraction des variantes qui n'impliquent pas de patrons syntaxiques prédéfinis. Cette approche dite « non-symbolique » consiste d'abord à extraire les groupes nominaux du corpus de manière automatique, et ensuite à les regrouper en séries de variantes en exploitant la similarité des chaînes de caractères.

Nous étudions les variantes terminologiques de type « composé morphologique - composé syntagmatique » dans une perspective multilingue, par conséquent nous ne voulons pas définir des patrons de variation en amont, ces patrons étant dépendants de la langue. Nous introduisons donc une méthode d'extraction des variantes de ce type qui sera indépendante de la langue, même si elle sera forcement moins précise qu'une méthode basée sur des patrons linguistiques prédéfinis. Nous souhaitons que cette méthode permette d'extraire, entre autres, des variantes avec expansion qui sont, à notre connaissance, peu étudiées jusqu'à présent.

8.2 Méthode d'identification des variantes

Parmi les travaux de l'état de l'art, nous trouvons deux scénarios d'extraction des variantes sans patrons linguistiques prédéfinis : l'approche non-symbolique décrite dans (Weller et al., 2011) exploitant la similarité de chaînes de caractères, et l'étape préliminaire de l'approche utilisée par Yoshikane et al. (1999) établissant des patrons morphosyntaxiques de variation. Dans ce travail, nous utilisons une méthode plus proche de (Yoshikane et al., 1999) à la différence que nous alignons les composés morphologiques extraits et segmentés par CompoST (à la place des termes polylexicaux d'une base de données) avec des termes composés syntagmatiques extraits du même corpus par TermSuite (à la place des phrases brutes extraites du corpus), cf. figure 8.1.

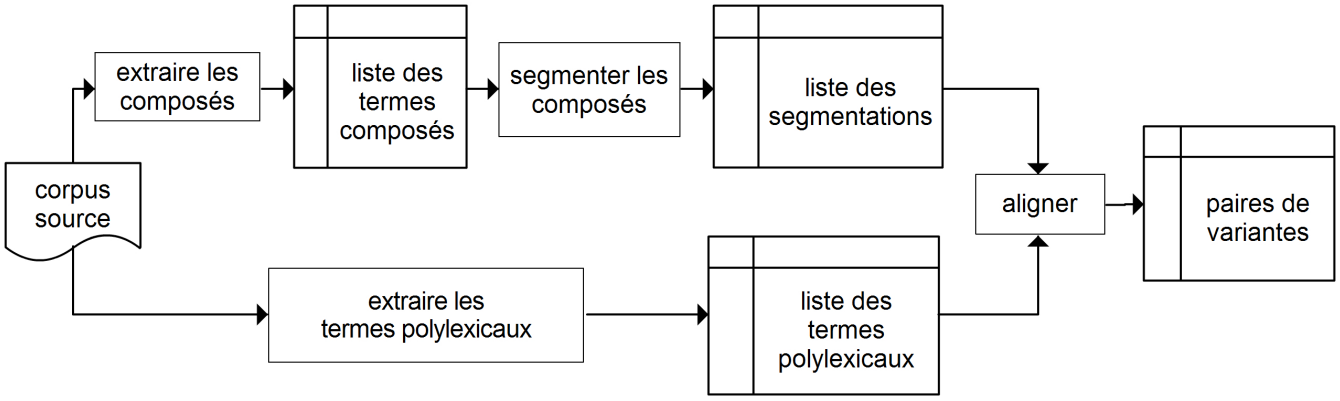

FIGURE 8.1 – De l'extraction des termes à l'alignement des variantes

L'alignement se fait si un CS contient les lemmes de tous les composants du CM. Nous autorisons une fenêtre de 15 caractères (y compris des espaces) entre les éléments du CM dans le CS pour couvrir les variantes quasi-synonymiques. Ce nombre correspond à la distance maximale entre les composants initiaux à l'intérieur du CS, attestée en DE pendant nos expériences préliminaires :

> DE Primärtumor 'tumeur primaire' ↔
> primäre intraokulare Tumor 'tumeur intraoculaire primaire'

L'ordre des éléments dans le CS peut être différent de l'ordre dans le CM :

> RU ветропоток ↔ поток ветра
> vetropotok ↔ potok vetra 'flux d'air';
>
> DE Häufigkeit-Ergebnis-Beziehung ↔ Beziehung von Häufigkeit und Ergebnis
> 'corrélation entre la fréquence et le résultat (du traitement)'.

Nous considérons donc toutes les permutations possibles des éléments.

L'alignement est indépendant de la langue, tandis que l'extraction des CM et CS comprend des règles linguistiques spécifiques à chaque langue. Il est plus facile et moins laborieux de définir des règles de formation des termes CS pour une langue (ces règles sont déjà formulées pour les langues que nous traitons dans TermSuite) que de définir des règles de variation, plus complexes car chaque type de CS peut avoir plusieurs variantes CM, et vice versa. Le fait d'utiliser la liste des termes extraits par TermSuite et non simplement le corpus brut permet d'assurer que les variantes CS détectées sont des termes bien formés et non juste des mots adjacents ne formant pas de syntagme.

Cette méthode est assez basique et ne prétend pas innover pour l'extraction des variantes. L'importance est accordée dans ce chapitre aux variantes extraites et à leurs structures.

8.3 Expériences et paramétrage des outils

Nous appliquons cette méthode aux corpus spécialisés DE, EN, FR, RU de deux domaines, *l'énergie éolienne* et le domaine médical : *cancer du sein* pour DE, EN, FR et *diabète* pour RU à défaut de corpus du cancer (pour les tailles des corpus, cf. annexe B).

Les corpus ont été lemmatisés et annotés avec les catégories grammaticales par TreeTagger. Pour appliquer CompoST, un lexique des lemmes a été construit à partir de chaque corpus. Ces lexiques contiennent seulement les noms et les adjectifs d'une longueur minimale de 6 caractères. Contrairement à nos expériences de segmentation et de traduction, dans le cas de l'extraction des variantes nous avons autorisé les

composés appartenant au dictionnaire bilingue. Par contre, nous avons exigé une fréquence minimale dans le corpus pour éviter de récupérer trop de candidats bruités. Cette fréquence minimale a été définie à 5 pour toutes les langues sauf FR pour lequel les premières expériences avec la fréquence de 5 ont amené un nombre dérisoire de composés et par conséquent de variantes. Nous avons donc diminué la fréquence minimale pour le français jusqu'à 2.

Comme nous l'avons constaté dans les expériences de segmentation, TreeTagger n'accomplit pas toujours correctement la lemmatisation des mots composés. Pour les lexiques DE, nous avons appliqué les mêmes règles de correction de la lemmatisation que nous utilisions dans nos expériences de segmentation (section 6.2.9). La correction de la lemmatisation pour d'autres langues n'a pas été effectuée.

Concernant le paramétrage de CompoST, nous avons utilisé le seuil optimisant le rappel car le but dans ces expériences n'est pas d'atteindre la précision la plus haute d'extraction, mais de découvrir des variantes des structures variées. Seule la segmentation la mieux classée par le système a été exploitée pour l'extraction des variantes. Pour ressources lexicales supplémentaires, nous avons utilisé les anti-dictionnaires et les listes de composants néoclassiques excluant les préfixes, car dans cette expérience nous ne cherchons pas à identifier les mots préfixés.

Pour l'extraction des termes par TermSuite, la fréquence minimale du CS a été définie à 2 pour écarter les hapax legomena et réduire le nombre d'erreurs. Nous n'avons pas augmenté cette valeur car une grande partie des variantes sont des occurrences rares et apparaissent occasionnellement dans les textes. Pour le français, nous avons fait une exception et utilisé une fréquence minimale de 1 pour augmenter le nombre de paires de termes extraites.

8.4 Résultats de l'extraction des variantes

Les expériences décrites ci-dessus ont abouti à l'extraction d'un certain nombre de candidats qui varie en fonction de la langue et du domaine : entre 23 (RU, médical) et 247 (EN, énergie éolienne). Ces chiffres indiquent le nombre de paires de variantes candidates et non le nombre d'occurrences de chaque variante dans le corpus, qui peut être bien plus élevé. Un écart important entre les nombres de candidats extraits s'explique par des proportions différentes de CM et de leurs variantes selon les langues et les domaines. Les corpus initiaux sont de taille comparables, sauf le corpus *diabète* pour le russe qui est légèrement moins étendu que les autres du domaine médical. Les résultats et quelques statistiques sont présentés dans le tableau 8.1.

TABLE 8.1 – Résultats de la détection des variantes « composé morphologique - composé syntagmatique »

	ÉNERGIE ÉOLIENNE				DOMAINE MÉDICAL			
	DE	EN	FR	RU	DE	EN	FR	RU
Taille de lexique après filtrage	3815	1617	2760	5970	3739	2602	4668	2510
Nb de CM candidats extraits	2281	456	793	1903	2092	888	1785	815
Nb de paires de variantes extraites	87	247	44	118	153	69	30	23
Nb de paires de variantes correctes	68	96	12	83	120	45	10	10
Dont les variantes graphiques	6	45	8	0	2	26	8	0
Dont les variantes empruntées	0	0	6	0	0	0	1	0
Nb d'erreurs dues à la segmentation	0	94	16	5	0	14	15	8
Précision (%)	78	39	27	70	78	65	33	43

8.4.1 Évaluation de la précision

Pour cette tâche, l'évaluation du rappel est très difficile à mettre en place car elle nécessite d'annoter préalablement toutes les variantes de toutes les structures possibles présentes dans les corpus. Nous nous sommes donc limitée à l'évaluation de la précision que nous calculons comme le nombre de paires de variantes extraites correctement parmi toutes les paires extraites :

$$P = \frac{\textit{nb paires de variantes correctes}}{\textit{nb paires de variantes extraites}} \tag{8.1}$$

Globalement, la quantité de paires de variantes extraites reste modeste par rapport au nombre de CM candidats proposés par CompoST, malgré le caractère permissif de la détection des variantes. Tous les candidats proposés en réalité ne sont pas des composés morphologiques : CompoST produit aussi des sur-segmentations, d'autant plus avec le seuil optimisant le rappel.

Analyse par langue. La langue la plus compositionnelle dans notre échantillon est l'allemand, et cela se traduit par un nombre relativement élevé de paires de variantes extraites, surtout pour le domaine médicale (153 paires vs. 87 pour l'énergie éolienne) et aussi par la précision la plus forte : 78 % pour les deux domaines. En anglais également un nombre élevé de variantes candidates ont été proposées (247 pour l'énergie éolienne et 69 pour le domaine médical), mais avec une précision moins élevée d'extraction (39 % et 65 % respectivement). En français, peu de variantes candidates ont été trouvées malgré la valeur de la fréquence minimale d'apparition des termes moins élevée que dans les autres langues : 44 pour l'énergie éolienne et 30 pour le domaine médical, avec une précision faible de 27 % et de 33 % respectivement et avec la présence d'emprunts de la langue anglaise parmi ces variantes (e.g. *windenergy* ↔ *wind energy*). En russe, une précision relativement haute a été obtenue pour l'énergie éolienne (70 % sur 118 candidats) contrairement à une précision bien moins élevée pour le domaine médical (43 % sur 23 candidats). En plus, en anglais et en français, le taux de variantes graphiques (*transistor-diode* ↔ *transistor diode*) est très élevé (jusqu'à 8 variantes graphiques sur 10 variantes correctes pour FR, domaine médicale), alors qu'en allemand leur pourcentage est très modeste (9 % pour l'énergie éolienne et 1,5 % pour le domaine médicale), et aucune variante graphique n'a été observée parmi les paires extraites en russe (ce type de variantes est possible en russe, mais pas dominante).

8.4.2 Analyse des erreurs

Nous avons analysé des alignements jugés incorrects. Parfois un des composants des CM et CS alignés a une sémantique différente :

EN over-speed 'excès de vitesse' ↔ overall speed 'vitesse moyenne';
RU парогенератор ↔ параметр генератора
parogenerator 'générateur de vapeur' ↔ parameter generatora 'paramètre du générateur'.

Dans d'autres cas jugés incorrects les composants du CM et du CS sont sémantiquement équivalents, mais reliés entre eux par des relations différentes :

DE Genveränderung ↔ Gen**test** auf Veränderungen
'mutation génétique' ↔ 'dépistage des mutations', lit. 'recherche génétique des mutations'.

L'attribut *génétique* se rapporte à des concepts différents au sein des deux termes. Un exemple d'alignement de ce type en FR est *cartographique* ↔ *carte **hydro**graphique*.

Un large nombre d'erreurs est lié à une lemmatisation impropre, surtout pour la langue russe avec ses six cas et donc un grand inventaire de formes fléchies. Ainsi, pour le corpus de l'énergie éolienne, parmi 35 paires jugées incorrectes, 16 sont en fait correctes, mais équivalentes à une autre paire déjà comptée comme correcte avec seulement la désinence qui change.

Certaines erreurs ont été introduites pendant l'extraction des CS. Parfois des termes incomplets sont extraits, par exemple DE *Risiko an Brustkrebs*, lit. 'risque + cancer du sein', partie du terme complexe *Risiko an Brustkrebs zu erkranken* 'risque de développer un cancer du sein'. Le candidat incomplet a ensuite été aligné avec le composé *Brustkrebsrisiko* 'risque de cancer du sein'.

Erreurs dues à la segmentation. Nous avons accordé une attention particulière aux erreurs provenant d'une segmentation incorrecte, notamment la sur-segmentation. Par exemple, le terme du corpus sur le cancer du sein EN *medication* 'médicaments' a été segmenté par CompoST en *medical* + *ion* et a ensuite été aligné avec les CS *medical condition* 'condition médicale', *medical examination* 'examen médical', *medical information* 'information médicale' et *medical opinion* 'opinion médicale'. Tous ces termes contiennent les sous-chaînes équivalentes aux composants *medical* et *ion* avec moins de 15 symboles d'écart entre eux. Pour le nombre d'erreurs dues à la sur-segmentation, voir le tableau 8.1. Ce problème de confusion entre un suffixe et un mot plein (cf. FR suffixe adjectival *-aire* vs. mot *aire*, etc.) a été nuisible dans l'extraction des variantes en EN (et particulièrement pour le corpus de l'énergie éolienne, 94 erreurs), en FR et en RU pour le domaine médical. Ces affixes « problématiques » n'ont pas été détectés au cours des expériences de segmentation car les mots les contenant s'avéraient généralement filtrés par leur présence dans les dictionnaires monolingues. Nous tenons à relever qu'aucune erreur de ce type n'a été attestée pour l'allemand : toutes les segmentations-candidates incorrectes ont été filtrées lors de l'étape d'alignement avec les CS. La conclusion en découle que pour EN, RU et FR, le paramétrage ou les ressources utilisées par CompoST doivent être révisés pour améliorer la qualité d'extraction des variantes avec la méthode proposée.

8.4.3 Structures des variantes

Nous avons examiné les paires de variantes correctes et les avons annotées avec les patrons morphosyntaxiques (du CM et du CS). Nous avons ensuite formulé les méta-patrons qui permettent de décrire la structure des variantes en faisant abstraction de la structure du syntagme particulière à une langue. Pour cela, les mots grammaticaux sont omis dans les méta-patrons. Les patrons avec des exemples pour chaque langue analysée sont présentés en annexe E. Un méta-patron peut avoir des réalisations morphosyntaxiques différentes selon la langue, mais aussi dans la même langue. Les méta-patrons sont définis à l'aide des symboles A, B, C, X, Y. A, B et C font référence aux composants du CM qui apparaissent également en CS. X et Y font référence aux éléments supplémentaires qui étendent le CS et qui ne sont pas présents en CM. Pour donner une idée de la productivité d'un patron, le nombre de paires de variantes extraites correspondant à chaque patron morphosyntaxique est indiqué. Les patrons les plus courants pour chaque langue sont mis en gras.

Voici deux méta-patrons attestés dans les quatre langues :

1. AB ↔ A B
   FR *microsystème - micro système*. Ce patron a été identifié pour les expressions syntagmatiques en anglais par Ville-Ometz et al. (2007). Parfois le composant gauche du CM correspond à une troncation d'un élément du CS : EN ***bio****diversity* ↔ ***biological*** *diversity*, 'biodiversité ↔ diversité biologique'. Ce phénomène nous fait nous poser la question : quelle forme dans une paire de variantes est originelle, parmi CM ou CS ? La tradition veut que dans une paire on analyse une forme comme étant celle de base, et l'autre comme variante. D'un côté, beaucoup de composés morphologiques sont formés étymologiquement à partir des composés syntagmatiques. D'un autre côté, pour des composés très spécialisés, la forme raccourcie (CM) est entrée dans les dictionnaires du domaine et semblent dominer dans les textes. Dans les expériences nous sommes partie des CM pour détecter les CS, mais ceci est un procédé lié seulement à la méthode proposée d'extraction des variantes. La forme de base devrait être désignée au cas par cas, par une analyse étymologique et une analyse d'usage des deux formes. Dans ce travail, une telle analyse n'a pas été effectuée, et nous nous permettons de garder le signe « ↔ » entre les deux formes.

2. AB ↔ B A
   En DE, EN et FR la partie CS de ce méta-patron est réalisée par la structure morphosyntaxique $N\ S{:}p\ N$ ($S{:}p$ signifie un article, une proposition ou un article contracté), contrairement au russe avec la structure $N\ N_{GEN}$:

   DE Mammakarzinom-Patientin ↔ Patientin mit Mammakarzinom
   'une patiente avec cancer du sein';

   EN blade-element ↔ element of blade 'élément de pale';

   FR hormonosensibilité ↔ sensibilité aux hormones;

   RU энергобаланс ↔ баланс энергии
   energobalans ↔ balans energii 'bilan d'énergie'.

Variantes avec expansion. Pour les deux méta-patrons, les composants présents dans le CM peuvent être étendus dans le CS avec un, voire deux autres éléments (X, Y) :

AB ↔ AX B
EN fixed-speed ↔ fixed-**rotational** speed
'à vitesse fixe' ↔ 'vitesse fixe de rotation';

AB ↔ B X A
RU энергоресурс ↔ ресурс ветровой энергии
energoresurs ↔ resurs **vetrovoj** energii
'ressource d'énergie' ↔ 'ressource d'énergie éolienne';

AB ↔ B X YA
DE Krebsrisiko ↔ Risiko **Erbliches Darm**krebs
'risque de cancer' ↔ 'risque héréditaire de cancer colorectal'

Les éléments insérés sont soit des lexèmes indépendants (et donc séparés par espace de leur gouverneur), soit concaténés à un des composants, soit séparés par un tiret ou parfois par un autre symbole ("/"). Cela dépend, bien sûr, de la langue, mais peut être très variable selon les textes, cf. les variantes du terme EN quasi-composé *low-speed* :

low wind speed
low wind-speed
low-wind speed.

Souvent la variante avec expansion désigne un hyponyme du CM, mais pas toujours. Quand le CM est formé par la réduction lexicale du CS, les deux formes sont synonymiques :

RU ветроресурс ↔ ветроэнергетический ресурс
vetroresurs ↔ vetro**ènergetičeskij** resurs
'ressource éolienne' ↔ 'ressource d'énergie éolienne'.

Lorsque le modificateur est étendue, les variantes conservent généralement la proximité conceptuelle (même en étant hyponymes). En revanche, quand la tête du syntagme subit une expansion, un certain écart conceptuel devient fréquent :

DE Stromerzeugung ↔ Erzeugungs**kosten** für Strom
'génération d'électricité' ↔ 'coût de génération d'électricité'.

Dans l'annexe E, les variantes de ce type sont en italique.

Les méta-patrons, dont le CM contient 3 éléments et les trois éléments correspondent aux éléments indépendants dans CS, ont été attestés dans nos expériences seulement pour l'allemand, et ils ne sont pas nombreux. Il s'agit soit d'une variante graphique :

Erneuerbare-Energien-Gesetz ↔ Erneuerbare Energien Gesetz
'loi sur l'énergie renouvelable',

soit d'une variante avec la coordination :

Häufigkeits-Ergebnis-Beziehung ↔ Beziehung von Häufigkeit und Ergebnis
'fréquence + résultat + rapport' ↔ 'rapport entre la fréquence et le résultat'.

Pour cette langue, plusieurs CM à trois et plus de trois composants ont été extraits, mais certains de leurs composants restent soudés après la transformation en CS :

Treibhausgasemission ↔ Emission von Treibhausgasen
'gaz à effet de serre + émission' ↔ 'émission de gaz à effet de serre'.

Pour décrire de tels cas, deux éléments de méta-patron A et B suffisent.

8.5 Bilan

Dans ce chapitre nous nous sommes intéressée aux variantes monolingues de type « composé morphologique - composé syntagmatique ». Notre but était de comparer la productivité et les structures de telles variantes dans les langues de nos expériences. Pour cela, nous avons appliqué une méthode générique d'extraction des variantes, basée sur la recherche des composants équivalents entre les CM et les CS extraits préalablement.

La précision obtenue varie entre 27 % (FR, énergie éolienne) et 78 % (DE, les deux domaines). La précision pour DE est legérèment plus élevée que celle obtenue par Weller et al. (2011) (74 %) dans une expérience semblable et moins élevée que celle obtenue par Yoshikane et al. (1999) (94 %) pour le japonais, dans les deux cas avec l'utilisation de patrons prédéfinis de variation. L'avantage est que nous n'avons pas utilisé de patrons qui seraient spécifiques à chaque langue.

L'analyse des variantes extraites a permis de formuler des méta-patrons de variation, d'estimer leur distribution dans les quatre langues et de décrire les patrons morphosyntaxiques correspondants. Les patrons décrits peuvent être utilisés ensuite pour une extraction plus précise dans chaque langue. La plus grande variété de patrons a été attestée pour l'allemand, qui est la langue la plus compositionnelle des quatre. En russe, un inventaire relativement large des variantes est aussi attesté, surtout pour le domaine de l'énergie éolienne. En français, au contraire, très peu de variantes ont été extraites pour les deux domaines, et avec une nette dominante des variantes graphiques. Cela s'explique par le fait que la composition n'est pas productive en français, mis à part le cas des composés néoclassiques qui tendent à former ce type des variantes. Pour une extraction des variantes en français, un corpus plus large et une technique plus avisée s'avèrent nécessaires. Un autre type de variantes que les composés néoclassiques peuvent donner sont les périphrases natives impliquant la substitution des formes supplétives « composant néoclassique » ↔ « radical natif » : par exemple, *annexectomie* ↔ *ablation des annexes de l'utérus* (*ectomie* ↔ *ablation*). Ce type de variation en français a été étudié par Grabar et Hamon (2014). La technique proposée dans ce chapitre ne permet pas d'extraire ce type de variantes. En anglais, les variantes graphiques sont aussi extrêmement présentes étant donné que dans cette langue les composants d'un CM sont souvent concaténés ou reliés par un tiret.

Parmi les variantes extraites il y a des variantes avec expansion (dont le CS est étendu par un ou deux éléments supplémentaires). L'expansion provoque souvent la spécification du sens du CS et parfois un écart conceptuel, mais cette règle n'est pas absolue et les exceptions sont possibles.

Ces expériences montrent également que la segmentation des composés morphologiques est utile pour l'extraction des variantes dans les langues dans lesquelles la composition est productive.

9

Conclusion

Synthèse

Ce travail a porté sur le traitement des termes composés dans le cadre de la construction des lexiques terminologiques multilingues. Étant donné leur forme graphique, les composés morphologiques sont généralement traités par les applications du TAL comme des termes simples, ce qui empêche de capturer leur complexité sémantique. L'identification et la segmentation des mots composés dans les langues propices à la formation des composés se sont déjà montrées utiles pour des tâches variées du TAL (traduction statistique, recherche d'information, etc.) qui dépassent largement le traitement des terminologies. Cependant, la composition est particulièrement productive dans les langues de spécialité, y compris dans des langues qui ne sont généralement pas considérées comme très compositionnelles (russe, français...). Par conséquent, le traitement particulier des composés est d'autant plus nécessaire dans les applications de la terminologie computationnelle.

Nous avons commencé le travail de cette thèse par une tâche terminographique qui consistait à compiler (de manière semi-manuelle) des **listes de référence de termes caractéristiques pour un domaine donné**, afin de les utiliser ensuite dans l'évaluation d'outils d'extraction automatique de termes. Nous avons construit les listes monolingues en français et en anglais, ainsi que les listes bilingues EN-FR, FR-RU pour deux domaines de spécialité : l'énergie éolienne et les technologies mobiles. La technique d'évaluation mettant en œuvre une liste de référence n'est pas sans défaut, car elle amène à comparer une liste de taille nécessairement limitée avec une liste de termes extraits par un outil à partir d'un corpus de textes potentiellement très longue. Néanmoins, cette technique permet de comparer à moindre coût les résultats de plusieurs systèmes.

A l'occasion de cette évaluation, nous avons comparé les erreurs commises par un extracteur automatique avec les difficultés qu'un terminologue peut rencontrer en construisant une liste de termes du domaine. Nous avons constaté que, dans les deux situations, le caractère graduel de spécificité d'une unité lexicale et l'interférence des domaines posent problème pour le choix des termes. Ces problèmes ne sont pas nouveaux, ils ont été révélés dans de nombreux travaux en terminologie computationnelle. L'extraction automatique a aussi ses faiblesses spécifiques, telles que la sensibilité à la taille des corpus utilisés et à leur degré de comparabilité, le découpage imparfait des termes polylexicaux (le problème du degré de l'unité, également bien connu dans les recherches terminologiques) et, pour l'extraction multilingue, le problème d'alignement de composés monolexicaux complexes d'une langue avec des composés polylexicaux d'une autre langue.

Le fait que les composés morphologiques se traduisent souvent vers d'autres langues par des composés syntagmatiques montrent le caractère artificiel de l'opposition entre les composés morphologiques et syntagmatiques dans le cadre multilingue. Dans le même temps, du point de vue de l'extraction monolingue, deux types de termes provoquent des difficultés différentes : les composés morphologiques doivent être segmentés en composants sémantiquement autonomes, tandis que les composés syntagmatiques doivent être reconnus en tant qu'unités. Dans cette thèse, nous nous sommes penchée sur ce premier défi.

Nous avons examiné des **méthodes de reconnaissance et de segmentation des mots composés**, à la recherche d'une méthode multilingue, adaptable au domaine de spécialité et à la langue, et capable de traiter l'allomorphie des composants. N'ayant pas trouvé de méthode qui correspondrait à tous ces critères, nous en avons proposé une nouvelle combinant certaines caractéristiques des autres approches, et qui a été implémentée dans l'outil « CompoST ».

CompoST s'applique à une liste de termes monolexicaux, préalablement lemmatisés, pour en retenir des termes composés (natifs, néoclassiques et d'autres) et aussi des termes préfixés, et de les mettre en correspondance avec les lemmes de leurs parties constituantes. Cette méthode est fondée sur l'utilisation du corpus, du dictionnaire et des similarités des chaînes de caractères, et de ce fait elle peut être utilisée de manière indépendante de la langue à condition d'avoir les ressources adéquates. Dans le même temps, elle peut être adaptée à une langue donnée, avec des règles linguistiques spécifiques et l'utilisation de ressources lexicales complémentaires. Elle est également adaptable au domaine de spécialité.

Nous avons testé cette méthode, ainsi que ses différentes composantes indépendamment, sur quatre langues (anglais, allemand, français et russe) et deux domaines (l'énergie éolienne et le cancer du sein), et nous avons obtenu des résultats compétitifs avec d'autres méthodes de l'état de l'art. Toutefois, notre méthode ne prétend pas résoudre l'ensemble des problèmes de la segmentation automatique : le traitement de l'allomorphie des composants reste à améliorer et la segmentation appropriée des termes parasynthétiques n'est pas systématique.

Nous avons démontré que la segmentation automatique des termes composés, même si elle n'est pas parfaite, est utile pour la construction des lexiques terminologiques, sur l'exemple de deux tâches : l'alignement lexical bilingue (la traduction) des termes composés et le regroupement des variantes.

L'alignement bilingue entre les termes composés et leurs équivalents polylexicaux peut se faire suivant la méthode compositionnelle de traduction. Nous avons appliqué cette méthode aux termes composés, identifiés et segmentés par CompoST. Les expériences ont été menées sur les mêmes domaines que celles de la segmentation, pour les couples de langues suivants : russe-anglais, anglais-français, allemand-anglais. Cela a permis de trouver des équivalents pour une partie des termes qui n'avaient pas de traduction dans les dictionnaires de langue générale, puis de générer des traductions polylexicales des termes monolexicaux. Le rappel et la précision de l'alignement ne sont pas très élevés, ce qui est dû non seulement à la qualité de la segmentation, mais aussi à la méthode compositionnelle d'alignement, entièrement fondée sur l'hypothèse de la compositionnalité des composés qui n'est pas toujours vraie. Pour donner une idée de la proportion des termes traduisibles compositionnellement pour les couples de langues étudiés, nous avons effectué l'alignement en utilisant la segmentation de référence. Les résultats obtenus indiquent la limite basse de la quantité de termes compositionnels, car une partie des erreurs provient de l'incomplétude des ressources utilisées. Les résultats varient selon la langue, mais la comparaison des rappels de la traduction obtenue en utilisant les segmentations produites par CompoST et celle obtenue en utilisant les segmentations de références fournit des valeurs comparables.

Un autre point d'application de la segmentation dans la construction des lexiques terminologiques, que nous avons pu aborder dans cette thèse, est **l'identification des variantes polylexicales (syntagmatiques) des termes composés**. Souvent les termes composés rivalisent dans les textes avec l'emploi de leurs variantes syntagmatiques. Pour les regrouper, nous avons appliqué une méthode basique fondée sur la recherche des composants des termes composés, identifiés par CompoST, dans une liste des termes polylexicaux extraits du même corpus. Nous avons choisi de ne pas définir de patrons de variation en amont

car ces patrons sont spécifiques à la langue. Cette méthode a permis de détecter des variantes de structures diverses, y compris des variantes peu étudiées avec l'expansion graphique. Ensuite, nous avons décrit des meta-patrons génériques de la variation et des patrons morphosyntaxiques pour chacune des langues de nos expériences. L'inconvénient de cette approche, assez brute, tient dans le grand nombre de candidats incorrects produits.

Limites du travail et perspectives

Nous indiquons les limites de nos travaux et à partir de celles-ci, proposons des perspectives d'amélioration.

Segmentation des termes composés. Nos expériences de segmentation des termes composés pour le russe ont montré que la composition est productive dans cette langue, surtout dans les textes spécialisés. La composition nécessite d'être traitée par des outils du TAL pour cette langue, ce qui n'est pas le cas actuellement, à notre connaissance. Cela ouvre une perspective de travail similaire pour d'autres langues slaves.

Notre méthode de segmentation se veut multilingue dans le sens où elle peut être utilisée pour une nouvelle langue sans avoir obligatoirement à définir des règles manuellement. Cependant, elle est limitée aux langues dans lesquelles le phénomène de la composition morphologique, tel que nous l'avons présenté, est courant. Actuellement, elle a été testée sur des langues indo-européennes, notamment des familles germanique, romane et slave. Cela serait intéressant de la tester également sur des langues ouraliennes. Il est possible que pour ces langues, caractérisées par l'agglutination, la segmentation en morphes soit plus utile que la segmentation en composants, et que les méthodes probabilistes dans l'esprit de *Morfessor* soient plus efficaces. Cependant, les méthodes de cette famille, en l'état, ne sont pas adaptées pour le traitement de l'allomorphie. Ce défaut sera peut-être corrigé à l'avenir d'une manière similaire à la solution mise en place dans *Allomorfessor* (Virpioja et al., 2009).

Le traitement de l'allomorphie des composants est le défi principal de la segmentation automatique dans les langues synthétiques. Dans CompoST, nous avons essayé de résoudre ce problème en utilisant des règles linguistiques ainsi que la similarité des chaînes de caractères, mais les règles doivent être formulées pour chaque nouvelle langue, et la similarité introduit parfois des erreurs.

Traduction compositionnelle (alignement bilingue). Nous avons utilisé la segmentation automatique pour traduire des termes composés monolexicaux qui correspondent aux termes polylexicaux dans une autre langue. Pour cela, nous avons appliqué une méthode de traduction compositionnelle relativement basique. Nous jugeons que la qualité des alignements ainsi obtenus est suffisante pour pouvoir les exploiter dans la construction de lexiques terminologiques bilingues de manière supervisée par un terminologue, et éventuellement dans un contexte de traduction compositionnelle en tant que composante de la traduction assistée par ordinateur. Pour la traduction statistique non-supervisée (sans intervention de traducteur), une version plus élaborée de cette approche devrait être appliquée (par exemple, celle utilisant des patrons de traduction ou la probabilité des traductions des composants). De plus, le classement des traductions candidates est indispensable.

Nous nous sommes limitée ici à l'alignement bilingue. Pour la construction des ressources terminologiques multilingues, des liens encore plus complexes entre les terminologies de plusieurs langues sont à établir.

Nous avons effectué l'alignement dans le sens d'une langue plus propice à la composition morphologique vers une langue moins compositionnelle. En poursuivant la piste de la traduction de vocabulaire spécialisé ou général, la traduction peut se faire également d'une langue moins compositionnelle vers une langue compositionnelle, voire entre deux langues compositionnelles.

Identification des variantes polylexicales des termes composés. Les patrons génériques et les patrons morphosyntaxiques de variation décrits dans ce travail peuvent être utilisés pour une extraction de variantes adaptée à chaque langue. Il serait utile de calculer la précision pour chaque patron séparément afin de définir les patrons plus et moins robustes dans chaque langue. Il serait aussi intéressant d'étudier s'il y a une corrélation entre le type de variation et le domaine de spécialité.

Dans ce travail, nous n'avons pas analysé les relations lexico-sémantiques entre le terme de base et la variante selon le patron de variation. Une telle étude pourrait servir pour la construction des bases de données terminologiques.

Les variantes identifiées au niveau monolingue peuvent aussi être exploitées, dans le cadre de la construction d'une ressource terminologique, pour établir des liens bilingues, notamment pour aligner un terme composé de la langue source avec un terme polylexical de la langue cible en traduisant la variante polylexicale du terme source vers la langue cible.

L'ensemble des travaux de cette thèse confirme que dans un contexte multilingue, la distinction entre les termes simples et complexes doit être nuancée. Les termes composés dits morphologiques, graphiquement simples, ont souvent des équivalents syntagmatiques dans d'autres langues. Dans le même temps, pour pouvoir établir cette équivalence, les termes composés doivent être identifiés et segmentés, c'est-à-dire que leurs composants doivent être alignés avec des unités lexicales autonomes. La segmentation sera bénéfique pour bien d'autres applications du TAL qui n'ont pas été étudiées dans cette thèse, notamment pour la recherche d'information mono- et multilingue, l'indexation, le résumé automatique, la reconnaissance de la parole, la construction des thésaurus et des ontologies, la conception des systèmes de question-réponse.

Ressources terminologiques et outils cités

Acabit Daille (1994)
taln09.blogspot.com/2009/03/acabit-acquisition-de-termes-partir-de.html
Extracteur de termes complexes permettant de lemmatiser un corpus étiqueté et d'évaluer le degré de spécificité des syntagmes du corpus afin de proposer à l'utilisateur une liste des candidats termes de ce corpus. Il propose également des variantes des termes extraits. L'approche est basée sur les règles linguistiques (analyse des patrons morphologiques), ainsi que sur les modèles statistiques (mesure d'association « log-likelihood »).

AntConc http://www.antlab.sci.waseda.ac.jp
Laurence Anthony (2011)
Concordancier qui permet de visualiser les occurrences de mots ou de groupes de mots dans leurs contextes, de lemmatiser des corpus bruts et de filtrer des mots grammaticaux (pour cela, des listes de lemmes et de mots grammaticaux sont requises), de classer les lemmes ou les formes fléchies par fréquence dans le corpus, d'obtenir des co-occurrences fréquentes, etc. Il sert également à comparer les fréquences dans un corpus donné avec les fréquences d'un corpus de référence pour retenir les mots-clés du corpus donné.

BananaSplit http://niels.drni.de/s9y/pages/bananasplit.html
Segmenteur de mots composés en allemand. L'outil utilise un dictionnaire et des règles de transformation de composants (incluses dans le programme). Chaque mot est segmenté en deux parties au maximum. Les mots appartenant au dictionnaire ne sont pas segmentés.

BLEU (Bilingual Evaluation Understudy) Papineni et al. (2002)
Système pour l'évaluation de la qualité de traduction automatique. Les traductions faites automatiquement sont comparées avec celles de traducteurs-experts, et les scores sont calculés. Le système exige des traductions de référence sous forme de phrases alignées.

DériF Namer (2003)
http://www.cnrtl.fr/outils/DeriF/
Analyseur morpho-sémantique pour le français qui décompose les mots morphologiquement construits (par dérivation, conversion, composition native ou néoclassique), reconstruit leur structure hiérarchique et génère une « pseudo-définition » du mot à partir de son analyse.

Grand Dictionnaire Terminologique http://www.gdt.oqlf.gouv.qc.ca
Banque terminologique de l'Office québécois de la langue française. Une fiche correspond à un concept lié à un domaine spécialisé, et regroupe les termes en français et en anglais et, parfois, dans d'autres langues (catalan, espagnol, galicien, italien, latin, portugais, roumain). La banque contient environ 10 000 fiches.

IMS Open Corpus Workbench (CWB) http://cwb.sourceforge.net
Outil de traitement des corpus de textes avec annotation linguistique (lemmatisation, étiquetage des catégories grammaticales). Il est employé pour de nombreuses opérations sur les corpus, notamment la construction des concordances pour un mot ou une expression, l'extraction des collocations, ou encore le classement du lexique du corpus selon les fréquences des mots.

LEXTER Bourigault (1994)
http://sira.u-bordeaux3.fr/ONLINE/IE10/bourigault.html
Extracteur de termes simples et complexes basé sur l'apprentissage des informations syntaxiques des textes. L'outil est également utilisé pour organiser l'ensemble des termes extraits en un « réseau terminologique ».

Morfessor http://www.cis.hut.fi/projects/morpho/
Analyseur morphologique indépendant de la langue, basé sur une méthode probabiliste d'acquisition d'un lexique de morphes (i.e. réalisations des morphèmes) d'une langue à partir de textes.

SMOR Schmid et al. (2004)
Analyseur morphologique pour l'allemand fondé sur une modélisation de grammaire (transducteur à états finis) et utilisant un lexique de morphèmes. L'outil permet de traiter la flexion, la dérivation et la composition en allemand.

Splitter for German compounds Weller et Heid (2012)
http://www.ims.uni-stuttgart.de/~wellermn/tools.html
Segmenteur de mots composés pour la langue allemande. L'outil nécessite un corpus monolingue annoté avec les catégories grammaticales. La segmentation en plusieurs parties est possible. Plusieurs segmentations candidates classées selon leur probabilité sont fournies à l'utilisateur avec les parties du discours des composants. La segmentation est possible pour les formes de base ainsi que pour les formes fléchies.

TERMIUM Plus http://termiumplus.gc.ca
Banque de données terminologiques et linguistiques du gouvernement du Canada. Elle propose aux utilisateurs des définitions de termes (majoritairement en anglais et en français, mais aussi en espagnol et en portugais) en précisant leur domaine, ainsi que plusieurs outils d'aide à la rédaction terminologique. La banque sert d'outil de normalisation canadien, elle comprend environ 4 millions de termes.

TermSuite http://code.google.com/p/ttc-project
Outil permettant l'extraction monolingue de termes simples et complexes à partir d'un corpus monolingue, ainsi que l'alignement des termes à partir de corpus bilingues avec les méthodes distributionnelle, compositionnelle et mixte. Les termes complexes polylexicaux sont définis dans cet outil comme des groupes nominaux contenant deux ou plusieurs mots pleins, et ils sont détectés par l'extracteur grâce aux patrons syntaxiques. Les termes simples sont détectés en s'appuyant sur leur catégorie grammaticale, ainsi que sur leur spécificité ou le nombre d'occurrences dans le corpus donné.

B

Ressources utilisées dans les expériences

B.1 Corpus et dictionnaires

Dans nos travaux de construction des listes de référence pour une évaluation d'extraction terminologique, nous avons utilisé des corpus comparables français et anglais relatifs aux domaines de *l'énergie éolienne* et des *technologies mobiles*. Pour la segmentation des termes composés puis la traduction compositionnelle, nous avons utilisé des corpus comparables français, anglais, allemand et russe du domaine de *l'énergie éolienne* (en version plus restreinte EN et FR que celle utilisée pour les RTLs) et des corpus comparables français, anglais et allemand du domaine du *cancer du sein*. Pour l'identification des variantes terminologiques, les mêmes corpus ont été utilisés, avec un corpus russe du domaine du *diabète* en plus. Les tableaux B.1 et B.2 récapitulent les tailles des corpus en nombre d'occurrences.

Corpus de l'énergie éolienne et des technologies mobiles sont construits dans le cadre du projet TTC et sont disponibles pour sept langues [1] , mais pour les expériences présentées dans cette thèse nous avons exploité seulement les parties anglaise, française, allemande et russe. Les textes ont été recueillis sur internet à l'aide de Babouk (Groc, 2011), un outil de construction automatique des corpus spécialisés, en utilisant une liste de mots-clés. Ces corpus existent en version étendue (utilisée pour le domaine de l'énergie éolienne pour la construction des RTLs) et restreinte (utilisée pour les autres expériences).

Corpus médicaux. Les corpus du *cancer du sein* sont disponibles pour trois langues : l'anglais, le français et l'allemand. Ces corpus contiennent des articles scientifiques publiés sur les sites web et les portails spécialisés et sélectionnés manuellement. Les articles sélectionnés répondent à deux critères : (1) ils appartiennent au discours scientifique, et (2) contiennent le mot-clé *cancer du sein* ou son équivalent dans d'autres langues. La partie allemande contient surtout des résumés d'articles scientifiques, ce qui explique qu'il est de taille plus petite que les autres. Le corpus comparable du *cancer du sein* n'existe pas en russe, c'est pourquoi pour les expériences d'identification des variantes terminologiques, nous avons utilisé un corpus russe d'un autre sous-domaine médical, *diabète et nutrition*. Ce corpus contient des articles scientifiques et vulgarisés relatifs à ce sujet et sélectionnés manuellement sur le web.

1. http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html.

TABLE B.1 – Taille des corpus utilisés pour la construction des listes terminologiques de référence

Domaine	EN	FR
Énergie éolienne	750,855	710,702
Technologies mobiles	308,263	302,634

TABLE B.2 – Taille des corpus utilisés pour la segmentation et l'identification des variantes

Domaine	DE	EN	FR	RU
Énergie éolienne	358,602	314,549	313,943	323,929
Médical	378,474	527,268	529,544	289,085

Liste de fréquences dans la langue générale. Pour calculer la spécificité d'un terme dans un domaine (cf. section 6.1.5), nous devons connaître sa fréquence dans la langue générale. Dans nos expériences, nous avons utilisé des listes de fréquences de lemmes issus d'un corpus général. Pour le russe, nous avons utilisé la liste de fréquences disponible librement, qui a été construite à partir du Corpus National Russe [2]. Pour l'anglais, le français et l'allemand nous avons extrait les listes de fréquences à partir de corpus d'actualités. Nous avons utilisé les corpus d'actualités comme des corpus généraux car ils ne sont pas spécifiques à des domaines techniques ou scientifiques, et ils sont disponibles pour plusieurs langues. Pour l'allemand nous avons exploité la partie monolingue du corpus parallèle DE-EN de la campagne d'évaluation de systèmes de traduction WMT 2008 [3]. Pour l'anglais et le français nous avons utilisé une sous-partie du corpus de la revue *New York Times* et de son homologue français *Le Monde*. Les collections d'articles ont été récupérées sur les catalogues LDC [4] pour l'anglais et ELRA [5] pour le français. Les corpus contiennent environ 10 million de mots chacun. Pour raison de facilité de traitement des données, nous avons utilisé des échantillons de ces corpus incluant des textes choisis de manière aléatoire sur l'année 2004. Les statistiques pour les corpus généraux et les listes de fréquences issus de ces corpus sont renseignées dans le tableau B.3.

Dictionnaires. Nous avons utilisé des dictionnaires bilingues EN-FR, DE-EN, RU-EN de langue générale pour les expériences en traduction compositionnelle des termes (cf. chapitre 7), et les parties monolingues des mêmes dictionnaires pour le filtrage des termes à segmenter (cf. chapitre 6). Le dictionnaire FR-EN a été obtenu sur le catalogue ELRA [6], tandis que pour les paires RU-EN et DE-EN *English-Russian full dictionary* [7] et *English-German dictionary* [8] ont été utilisées. Le tableau B.4 récapitule le nombre d'entrées dans chaque dictionnaire bilingue.

B.2 Ressources lexicales pour la segmentation et la traduction

Pour valider ou non les composants candidats, CompoST effectue une recherche dans deux lexiques : le premier est issu du corpus de textes, le deuxième (optionnel) est construit à partir du dictionnaire et peut être enrichi par une liste d'éléments néoclassiques et de préfixes. En plus, un anti-dictionnaire peut être utilisé.

2. http://corpus.leeds.ac.uk/serge/frqlist/rnc-modern-lpos.num.html
3. http://www.statmt.org/wmt13/
4. http://catalog.ldc.upenn.edu/LDC2008T19
5. http://catalog.elra.info/product_info.php?products_id=438&language=fr
6. http://catalog.elra.info/product_info.php?products_id=667
7. http://dicto.org.ru/xdxf.html
8. http://www1.dict.cc/

TABLE B.3 – Listes de fréquences de la langue générale et corpus généraux

	DE	EN	FR	RU
Corpus général	104,908,852	5,001,609	5,001,974	109,115,810
Liste de fréquences	541,675	112,247	94,524	27,261

TABLE B.4 – Taille des dictionnaires bilingues

Paire de langues	EN-FR	DE-EN	RU-EN
Nb. d'entrées	145,542	777,174	526,876

Lexique issu du corpus. CompoST est conçu pour la segmentation des termes. Il prévoit donc l'utilisation d'une liste de spécificité des mots du corpus relatif à un domaine (sur le calcul de spécificité, cf. section 6.1.5). Pour chaque langue et domaine, cette liste de spécificité a été construite en s'appuyant sur la fréquence des mots dans le corpus spécialisé et dans un corpus général (décrits dans la section B.1). Les corpus avaient été préalablement lemmatisés et étiquetés par catégorie grammaticale. Le lexique obtenu ne contient donc que des lemmes, et il a été filtré par catégorie grammaticale. Pour ce travail, nous avons inclus uniquement les noms, adjectifs, verbes et adverbes dans les lexiques pour diminuer les erreurs possibles. Nous sommes consciente que les composants des composés peuvent appartenir à une autre catégorie grammaticale (pronoms, adjectifs numéraux), mais dans les langues traitées ces formations sont minoritaires. Le lexique final a été également filtré selon la fréquence : les mots avec une fréquence absolue dans le corpus spécialisé inférieure à 5 ont été écartés afin d'éviter les « non-mots ».

CompoST peut aussi être utilisé pour la segmentation des composés issus de la langue générale. Dans ce cas, la liste de spécificité n'a pas de sens et elle doit être remplacée par une liste de fréquences relatives d'un corpus général. Cette variante d'emploi a été appliquée pour l'expérience décrite dans la section 6.2.6, et le lexique utilisé pour ce but avait été également filtré par catégorie grammaticale et fréquence.

Lexique issu du dictionnaire. Afin d'obtenir ce lexique pour chaque langue, le dictionnaire monolingue a été filtré par catégorie grammaticale : comme pour le lexique issu du corpus, uniquement les noms, adjectifs, verbes et adverbes ont été sélectionnés.

Liste des éléments néoclassiques et des préfixes. Pour les expériences en traduction compositionnelle (chapitre 7), nous avons utilisé des NCP-listes bilingues. Ces listes contiennent des racines d'origine latine ou grecque, alignées avec leurs traductions, ainsi que certains préfixes, également avec leurs traductions. Pour obtenir cette ressource, nous avons combiné les listes d'éléments néoclassiques EN-FR et FR-DE, élaborées (à partir d'une liste de (Béchade, 1992) pour le français, qui a ensuite été traduite vers EN et DE) et complétées de manière semi-automatique par Harastani (2014), avec les tables de traduction de morphèmes DE-EN et EN-FR, constituées par Estelle Delpech et disponibles sur internet[9]. Pour obtenir la liste RU-EN, nous avons traduit la partie anglaise de la liste EN-FR vers le russe (quand la traduction néoclassique d'une racine existait). Les tables de traduction de morphèmes contenaient non seulement des racines néoclassiques, mais aussi des préfixes et des suffixes. Nous avons écarté les suffixes; en revanche, les préfixes ont été conservés. Un préfixe d'origine latine ou grecque peut être traduit vers une autre langue soit également par un préfixe néoclassique, soit par un préfixe natif (EN *mono-* ↔ RU одно-). Nous avons inclus les préfixes de deux types dans les listes. Une entrée de liste bilingue est un élément mis en correspondance avec une traduction. Lorsqu'un élément de la langue source a plusieurs traductions vers la langue cible, des entrées différentes sont créées. Cela explique une différence importante entre les tailles des listes EN-FR et

9. http://www.lina.univ-nantes.fr/?Linguistic-resources-from-the,1676.html.

TABLE B.5 – Taille des listes bilingues d'éléments néoclassiques et de préfixes

Paire de langues	EN-FR	DE-EN	RU-EN
Nb. d'entrées	891	481	143

TABLE B.6 – Taille des anti-dictionnaires

Langue	DE	EN	FR	RU
Nb. d'entrées	31	9	5	11

RU-EN, même si à la base nous sommes partie de la même liste EN. Les tailles des NCP-listes bilingues sont récupérées dans le tableau B.5). Pour les expériences de segmentation (chapitre 6), seules les parties monolingues ont été utilisées.

Anti-dictionnaires. Pour diminuer le nombre de fausses segmentations, CompoST prévoit l'utilisation optionnelle d'un anti-dictionnaire. Pour nos expériences, nous avons compilé des anti-dictionnnaires de taille modeste (entre 5 et 31 éléments, cf. tableau B.6), en nous basant sur des expériences préliminaires. Ils contiennent des mots grammaticaux mal identifiés par l'étiqueteur de catégorie grammatical ou mal lemmatisés, et certains mots pleins (lemmatisés par l'étiqueteur) qui ne forment généralement pas des composés, mais qui sont fréquents en tant que sous-chaîne des mots non-composés. Voici quelques exemples : FR *que, age*, EN *lay, ent*.

C

Paramètres de CompoST retenus selon la configuration

Configuration	Coefficients (α β γ δ)	Seuil	
		rappel optimal	précision optimale
DE Cancer du sein			
BASE	0,4 0,4 0,1 0,1	0,7	0,8
BASE+NCP+STOP	0,2 0,4 0,1 0,3	0,6	0,65
BASE+RÈGLES	0,8 0,1 0,1 0,0	0,85	0,9
BASE+NCP+STOP+RÈGLES	0,5 0,3 0,1 0,1	0,8	0,85
Fréquence générale	0,6 0,3 0,1 0,0	0,85	0,95
EN Énergie éolienne			
BASE	0,6 0,2 0,1 0,1	0,8	0,85
BASE+NCP+STOP	0,6 0,2 0,1 0,1	0,8	0,85
BASE+RÈGLES	0,7 0,1 0,1 0,1	0,8	0,85
BASE+NCP+STOP+RÈGLES	0,7 0,1 0,1 0,1	0,8	0,85
Fréquence générale	0,7 0,3 0,0 0,0	0,85	0,9
FR Cancer du sein			
BASE	0,5 0,1 0,1 0,3	0,6	0,65
BASE+NCP+STOP	0,7 0,1 0,2 0,0	0,85	0,85
BASE+RÈGLES	0,5 0,1 0,1 0,3	0,6	0,65
BASE+NCP+STOP+RÈGLES	0,5 0,1 0,1 0,3	0,6	0,7
Fréquence générale	0,5 0,1 0,0 0,4	0,55	0,6
RU Énergie éolienne			
BASE	0,3 0,1 0,1 0,5	0,4	0,45
BASE+NCP+STOP	0,3 0,1 0,1 0,5	0,4	0,45
BASE+RÈGLES1	0,3 0,1 0,1 0,5	0,4	0,45
BASE+RÈGLES2	0,5 0,1 0,3 0,1	0,8	0,85
BASE+NCP+STOP+RÈGLES2	0,3 0,1 0,4 0,2	0,7	0,8
Fréquence générale	0,8 0,1 0,1 0,0	0,9	0,95

D

Extraits de segmentation

Domaine : énergie éolienne.
Configuration de CompoST : BASE+NCP+STOP+RÈGLES avec le seuil optimisé pour la précision.
Pour les non-composés, l'indication « *no split* » est précisée. Tous les mots sont en minuscules, y compris les noms allemands, pour faciliter le traitement.

Candidat	Segmentation de référence	Segmentation CompoST
	DE	
abstandsregelungen	abstand regelung	abstand regelung ab stand regelung abs tand regelung
regelungsabsicht	regelung absicht	regelung absicht
pumpspeicherkraftwerken	pumpspeicher kraftwerk pumpe speicher kraft werk pumpe speicher kraftwerk	pumpspeicher kraftwerk
hauptstranges	haupt strang	hauptstrom ges haupt strang
beschleunigt	no split	no split
erfordert	no split	no split
energieaußenpolitik	energie außenpolitik energie außen politik	energie außenpolitik
production	no split	no split
spezif	no split	no split
tonhaltigkeiten	ton haltigkeit	no split
wicklungen	no split	wicklung gen
asynchronmaschinen	asynchron maschine	asynchron maschine
windpower	wind power	wind power winde power
immissionsaufpunkten	immission aufpunkt immission auf punkt	immission saufen punkt immission aufpunkt
dena-netzstudie	no split	no split

Candidat	Segmentation de référence	Segmentation CompoST
fehlerfolge	fehler folge	fehler folge fehlen erfolg fehler folgen fehlen folge fehl folge
geführt	no split	no split
	EN	
labtech	no split	no split
airfoil	air foil	air foil
evidence-based	evidence base	evidence base
stellenbosch	no split	no split
agreements	no split	no split
mcgraw-hill	no split	no split
signed	no split	no split
colebatch;	no split	no split
stall-regulated	stall regulate	stall regulate stall regular
interrelationship	inter relationship	inter relationship in ter relationship interpretation ship interrelation ship
geared	no split	no split
problems	no split	no split
reynolds	no split	no split
speech-language-hearing	speech language hearing	speech language hearing
interactions	inter action	inter action in ter action
www.embracewind.com	no split	no split
twenty-year	twenty year	twenty year
pgenerator	no split	no split
full-span	full span	full span
ridgeville	no split	no split
waterbird	water bird	water bird
four-point	four point	four point
standalone	stand alone	stand alone stand one
fixed-speed	fix speed	fix speed fixed speed
masters	no split	no split
taivalkoski	no split	no split
up-wind	up wind	no split
cash-back	cash back	cash back
grid-side	grid side	grid side
augmented	no split	no split
freiburg	no split	no split
transverse-flux	transverse flux	transverse flux
impermissible	im permissible	no split
networks	no split	net work net works

Candidat	Segmentation de référence	Segmentation CompoST
	FR	
puissance-vitesse	puissance vitesse	puissance vitesse
inverters	no split	no split
diesel-éoliens	diesel éolien	diesel éolien diesel éolienne
plaisance	no split	no split
jusqu’	no split	no split
sous-optimale	sous optimal	sous optimal
dispersed	no split	no split
coolmos	no split	no split
simulinktm	no split	no split
proven	no split	no split
plomb-acide	plomb acide	no split
product	no split	no split
project	no split	no split
transistor-diode	transistor diode	transistor diode transiter diode
integrated	no split	no split
optimalité	no split	no split
profile	no split	pro fil
maxima	no split	no split
survitesse	sur vitesse	sur vitesse
comporte	no split	no split
grammien	no split	no split
multivariable	multi variable	multi variable
super-conducteur	super conducteur	super conducteur
multi-pale	multi pale	multi pale
cadre-là	no split	no split
currents	no split	no split
energie	no split	no split
stator-rotor	stator rotor	stator rotor station rotor statère rotor stature rotor statuer rotor
blade-pitch	no split	no split
mono-pales	mono pale	mono pale
potential	no split	no split
stockable	no split	no split
extended	no split	no split
pré-magnétisation	pré magnétisation	pré magnétisation
dévolteur	no split	no split
surfacique	no split	no split
switching	no split	no split
multi-étages	multi étage	multi étage
évidement	no split	no split
diveux	no split	no split
supercondensateur	super condensateur	super condensateur
demi-cylindres	demi cylindre	demi cylindre

Candidat	Segmentation de référence	Segmentation CompoST
	RU	
гелиостанция	гелио станция	гелио станция
каскадный	no split	no split
ландау	no split	no split
киловатт-часов	киловатт час кило ватт час	no split
гидроаккумулирующий	гидро аккумулирующий гидро аккумулировать	no split
кыштымский	no split	no split
тонкоплёночные	тонкий плёночный тонкий плёнка	no split
курчатовский	no split	no split
интернет-магазин	интернет магазин	интер нет магазин
определённого	no split	no split
концентрирование	no split	no split
в-установок	no split	no split
ресурсный	no split	no split
теплотворный	no split	no split
аэростатный	no split	no split
федоров	no split	no split
омский	no split	no split
электрочайник	электро чайник	электро чайник
высокопроизводительный	высоко производительный высокий производительный	высокий производитель высоко производитель высокий производительный высоко производительный
максим	no split	no split
ретранслятор	ре транслятор	no split
василий	no split	no split
однотипный	один тип	одно тип
гуревич	no split	no split
повсеместный	no split	no split
слежение	no split	no split
конкурентоспособность	конкурент способность	конкурент способность
перспективность	no split	no split
фотокатализ	фото катализ	фото катализ
ракетный	no split	no split
латенто	no split	no split
высокоскоростный	высоко скоростной высокий скоростной высокий скорость высоко скорость	no split
сверхдешёвых	сверх дешёвый	no split
свс-амур	no split	no split
вышеперечисленный	выше перечислить выше перечисленный	no split
водосброс	вода сброс	вода сброс

E

Patrons de la variation

Notations utilisées :
E signifie un élément du composé (au sein du composé, la catégorie grammaticale ne peut pas toujours être définie),
$S{:}p$ - un article, une proposition ou un article contracté,
ADJ signifie dans ce tableau un adjectif ou un participe du verbe,
Z - un mot plein (souvent N ou ADJ, ou parfois un élément néoclassique) qui ne subit pas de modifications dans la variante, dans le cas de variantes graphiques.
L'index indique le lien morphologique entre les éléments.
WE indique le nombre de variantes du type correspondant attestées dans le domaine de l'énergie éolienne (« *wind energy* »),
MED - dans le domaine médical.
Les patrons les plus fréquents pour chaque langue sont mis en gras, les patrons qui engendrent un écart conceptuel sont en italique. Les variantes empruntées, attestées en français, ne sont pas prises en compte dans ce tableau.

Méta-patron	Patron morphosyntaxique	Exemple	Traduction FR	WE	MED
		DE			
AB ↔ A B	**Z_1Z_2 ↔ Z_1 Z_2**	**Windturbine ↔ Wind Turbine**	**turbine éolienne**	**11**	**6**
AB ↔ A B (avec troncation)	**E_1N_2 ↔ ADJ_1 N_2**	**Magnetfeld ↔ magnetische Feld**	**champ magnétique**	**3**	**21**
AB ↔ AX B	E_1N_2 ↔ E_1ADJ_3 N_2	Brustkrebsmortalität ↔ brustkrebsbedingte Mortalität	lit. cancer du sein + mortalité ↔ mortalité liée au cancer du sein	4	3
AB ↔ A X B	E_1N_2 ↔ ADJ_1 ADJ_3 N_2	Primärtherapie ↔ primäre endokrine Therapie	thérapie primaire ↔ hormonothérapie primaire	2	4
AB ↔ B A	**E_1N_2 ↔ N_2 S:p N_1**	**Elektrizitätserzeugung ↔ Erzeugung von Elektrizität**	**production d'électricité**	**22**	**49**

Méta-patron	Patron morphosyntaxique	Exemple	Traduction FR	WE	MED
	$E_1N_2 \leftrightarrow ADJ_2\ N_1$	Therapieerfolg ↔ erfolgreiche Therapie	succès de la thérapie ↔ thérapie fructueuse	0	1
	$E_1ADJ_2 \leftrightarrow ADJ_2\ N_1$	windstark ↔ starker Wind	venteux ↔ vent fort	1	3
AB ↔ BX A	*$E_1N_2 \leftrightarrow E_2N_3$ S:p N_1*	*Stromerzeugung ↔ Erzeugungskosten für Strom*	*génération d'électricité ↔ coût de génération d'électricité*	1	2
	$E_1N_2 \leftrightarrow E_2ADJ_3\ N_1$	*Stromerzeugungskosten ↔ kostenoptimale Stromerzeugung*	*coût de génération d'électricité ↔ génération d'électricité au meilleur coût*	1	0
AB ↔ B XA	**$E_1N_2 \leftrightarrow N_2$ S:p E_3N_1**	**Effizienzsteigerung ↔ Steigerung der Energieeffizienz**	**hausse d'efficacité ↔ hausse d'efficacité énergétique**	**20**	**25**
AB ↔ B X A	$E_1N_2 \leftrightarrow N_2$ ADJ:GEN_3 N:GEN_1	Energienutzung ↔ Nutzung erneuerbarer Energien	utilisation d'énergie ↔ utilisation d'énergie renouvelable	2	2
	$E_1N_2 \leftrightarrow ADJ_2\ ADJ_3\ N_1$	Lymphknotenbefall ↔ befallene axilläre Lymphknoten	atteinte ganglionnaire, lit. ganglions + atteinte ↔ ganglions axillaires atteints	0	1
	$E_1N_2 \leftrightarrow N_2$ S:p ADJ_3 N_1	Mammakarzinompatientin ↔ Patientin mit kleinen Mammakarzinomen	patiente atteinte du cancer du sein, lit. carcinome mammaire + patiente ↔ patiente avec un petit carcinome mammaire	0	2
AB ↔ B X YA	$E_1N_2 \leftrightarrow N_2$ ADJ:GEN_3 E_4N:GEN_1	Krebsrisiko ↔ Risiko Erbliches Darmkrebs	risque de cancer ↔ risque héréditaire de cancer colorectal	0	1
ABC ↔ C A B	$E_1E_2N_3 \leftrightarrow N_3$ S:p N_1 CONJ N_2	Häufigkeits-Ergebnis-Beziehung ↔ Beziehung von Häufigkeit und Ergebnis	lit. fréquence + résultat + rapport ↔ rapport entre la fréquence et le résultat	0	1
ABC ↔ A B C	$Z_1Z_2Z_3 \leftrightarrow Z_1\ Z_2\ Z_3$	Erneuerbare-Energien-Gesetz ↔ erneuerbare Energien Gesetz	lit. renouvelables + énergies + loi ↔ loi sur les énergies renouvelables	1	0
		EN			
AB ↔ A B	**$Z_1Z_2 \leftrightarrow Z_1\ Z_2$**	**blade-element ↔ blade element**	**élément d'une pale**	**46**	**28**
	$E_1ADJ_2 \leftrightarrow ADJ_1\ N_2$	right-sided ↔ right side	(du) côté droit	0	1
	$E_1ADJ_2 \leftrightarrow N_1\ N_2$	*grid-connected ↔ grid connection*	*raccordé au réseau ↔ connexion au réseau*	4	0
AB ↔ A B (avec troncation)	$E_1N_2 \leftrightarrow ADJ_1\ N_2$	biodiversity ↔ biological diversity	biodiversité ↔ diversité biologique	1	0
AB ↔ AX B	$E_1N_2 \leftrightarrow E_1ADJ_3\ N_2$	fixed-speed ↔ fixed-rotational speed	à vitesse fixe ↔ vitesse fixe de rotation	6	3
	$E_1N_2 \leftrightarrow E_1N_3$ S:p N_2	*airflow ↔ airfoil with flow*	*flux d'air ↔ profil aérodynamique avec le flux*	1	0
AB ↔ A XB	*$E_1ADJ_2 \leftrightarrow N_1\ E_3N_2$*	*grid-connected ↔ grid interconnection*	*raccordé au réseau ↔ interconnexion des réseaux*	3	0
AB ↔ A X B	$E_1N_2 \leftrightarrow N_1\ N_3\ N_2$	air flow ↔ air mass flow	flux d'air ↔ flux de la masse d'air	6	0

Méta-patron	Patron morphosyntaxique	Exemple	Traduction FR	WE	MED
	$E_1N_2 \leftrightarrow ADJ_1\ ADJ_3\ N_2$	high-risk ↔ high familial risk	à risque élevé ↔ risque familial élevé	7	4
	$E_1N_2 \leftrightarrow ADJ_1\ N_3\ N_2$	high-performance ↔ high plant performance	de haute performance ↔ bon rendement de la centrale	11	0
	$E_1N_2 \leftrightarrow ADJ_1\ ADV_3\ N_2$	large-scale ↔ large enough scale	à grande échelle ↔ à une échelle suffisamment grande	1	0
	$E_1N_2 \leftrightarrow ADJ_1\ CONJ\ ADJ_3\ N_2$	horizontal-axis ↔ horizontal and vertical axis	à axe horizontal ↔ axe horizontal et vertical	2	2
	$E_1ADJ_2 \leftrightarrow ADJ_1\ CONJ\ ADJ_3\ N_2$	left-sided ↔ left and right sides	du côté gauche ↔ côtés gauche et droit	0	1
	$E_1N_2 \leftrightarrow E_1\ CONJ\ ADJ_3\ N_2$	hydropower ↔ hydro and nuclear power	énergie hydraulique ↔ énergie hydraulique et nucléaire	1	0
AB ↔ B A	$E_1N_2 \leftrightarrow N_2\ S{:}p\ N_1$	blade-element ↔ element of blade	élément de pale	3	1
	$E_1ADJ_2 \leftrightarrow ADJ_2\ N_1$	receptor-positive ↔ positive receptor	récepteur positif	3	2
AB ↔ BX A	*$E_1N_2 \leftrightarrow N_2\ S{:}p\ N_1$*	*landowner ↔ ownership of land*	*propriétaire des terres ↔ propriété des terres*	1	0
AB ↔ B X A	$E_1ADJ_2 \leftrightarrow ADJ_2\ ADJ_3\ N_1$	node-positive ↔ positive axillary nodes	lit. ganglion + positif ↔ ganglions axillaires positifs	0	2
	$E_1ADJ_2 \leftrightarrow ADJ_2\ N_3\ N_1$	receptor-negative ↔ negative oestrogen receptor	récepteur négatif ↔ récepteur œstrogénique négatif	0	1
		FR			
AB ↔ A B	**$Z_1Z_2 \leftrightarrow Z_1\ Z_2$**	**microsystème ↔ micro système**		**6**	**7**
AB ↔ A B (avec troncation)	$E_1N_2 \leftrightarrow N_1\ ADJ_2$	oncogénétique ↔ oncologie génétique		0	1
AB ↔ B A	$E_1N_2 \leftrightarrow N_2\ S{:}p\ N_1$	hormonosensibilité ↔ sensibilité aux hormones		0	1
		RU			
AB ↔ A B (avec troncation)	**$E_1N_2 \leftrightarrow ADJ_1\ N_2$**	диетотерапия ↔ диетическая терапия	**thérapie diététique**	**38**	**4**
AB ↔ AX B	**$E_1N_2 \leftrightarrow E_1ADJ_3\ N_2$**	ветроресурс ↔ ветроэнергетический ресурс	**ressource éolienne ↔ ressource d'énergie éolienne**	**24**	**0**
AB ↔ A X B	*$E_1N_2 \leftrightarrow N_1\ S{:}p\ ADJ_3\ N_2$*	диетотерапия ↔ диета в комплексной терапии	*thérapie diététique ↔ diète dans un traitement d'association (lit. diète dans une thérapie d'association)*	0	1
	$E_1N_2 \leftrightarrow N_1\ ADJ{:}GEN_3\ N{:}GEN_2$	энергоресурс ↔ энергия водных ресурсов	*ressource énergétique ↔ énergie des ressources de l'eau*	2	0
	$E_1N_2 \leftrightarrow ADJ_1\ N_3\ N{:}GEN_2$	нормогликемия ↔ нормальный уровень гликемии	normoglycémie ↔ niveau normal de la glycémie	0	1
AB ↔ B A	**$E_1N_2 \leftrightarrow N_2\ N{:}GEN_1$**	энергобаланс ↔ баланс энергии	**lit. énergie + bilan ↔ bilan d'énergie**	**10**	**3**

Méta-patron	Patron morphosyntaxique	Exemple	Traduction FR	WE	MED
	$E_1 ADJ_2 \leftrightarrow ADJ_2\ N_1$	пароводяной ↔ водяной пар	*à la vapeur d'eau (lit. vapeur + eau)* ↔ *vapeur d'eau*	1	0
	$E_1 N_2 \leftrightarrow N_2$ S:p N_1	инсулинорезистентность ↔ резистентность к инсулину	insulinorésistance ↔ résistance à l'insuline	0	1
AB ↔ B X A	$E_1 N_2 \leftrightarrow N_2$ N:GEN$_3$ N:GEN$_1$	водоподъем ↔ подъем уровня воды	montée d'eau ↔ montée du niveau d'eau	3	0
	$E_1 N_2 \leftrightarrow N_2$ ADJ:GEN$_3$ N:GEN$_1$	энергопотребитель ↔ потребитель электрической энергии	consommateur d'énergie ↔ consommateur d'énergie électrique	5	0

Bibliographie

Peter Ackema et Ad Neeleman. The role of syntax and morphology in compounding. Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 21–37. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 63

AFNOR. *Document Afnor X 03-003, norme ISO 1087 : Terminologie, Principes et coordination*. Afnor, 1990. 23

Khurshid Ahmad, Andrea Davies, Heather Fulford, et Margaret Rogers. What is a term? The semi-automatic extraction of terms from text. Dans *Translation Studies: An Interdiscipline*, pages 267–278, Amsterdam/Philadelphia, 1992. John Benjamins. 38, 81, 94

Enrique Alfonseca, Slaven Bilac, et Stefan Paries. German decompounding in a difficult corpus. Dans *Proceedings of CICLING 2008*, pages 128–139, 2008. 59

Dany Amiot et Georgette Dal. La composition néoclassique en français et l'ordre des constituants. *La composition dans les langues, Artois Presses Université*, pages 89–113, 2008. 64, 84

Mark Aronoff. *Morphology by Itself: Stems and Inflectional Classes*. Linguistic Inquiry Monographs. The MIT Press, 1993. 63

Timothy Baldwin et Takaaki Tanaka. Translation by Machine of Complex Nominals: Getting it Right. Dans *Proceedings of the ACL'04 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, 2004. 108, 109, 110

Charles Bally. *Linguistique générale et linguistique française*. Francke, Berne, 1965. 62

Laurie Bauer. *English Word-formation*. Cambridge University Press, Cambridge, 1983. 61, 62, 64, 65, 83

Hervé-D. Béchade. *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France, 1992. 64, 65, 84, 141

Emile Benveniste. *Problèmes de linguistique générale*. Gallimard, Paris, 1974. 62, 63, 69, 80

Geert Booij. *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford Textbooks in Linguistics. Oxford University Press, 2005. 68

Didier Bourigault. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse, Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994. 41, 136

Didier Bourigault, Isabelle Gonzalez-Mullier, et Cécile Gros. Lexter, a natural language processing tool for terminology extraction. Dans *Proceedings of the 7th EURALEX International Congress*, pages 771–779, Goteborg, Sweden, 1996. 41

Didier Bourigault et Monique Slodzian. Pour une terminologie textuelle. *Terminologies Nouvelles*, 19: 29–32, 1999. 22

R. Boutin-Quesnel, N. Belanger, N. Kerpan, et L.-J. Rousseau. *Vocabulaire systèmatique de la terminologie*. Les publications du Québec (Les cahiers de l'Office québécois de la langue française). OLF, Québec, 1985. 21

Lynne Bowker et Jennifer Pearson. *Working with specialized language: a practical guide to using corpora*. Routledge, London, 2002. 24, 25, 27

M. Braschler et B. Ripplinger. How effective is stemming and decompounding for german text retrieval. Dans *Information Retrieval*, pages 291–316, 2004. 59

Hadumond Bussmann. *Routledge Dictionary of Language and Linguistics*. Routledge, London, 1996. 61, 64

Maria Térésa Cabré. *La terminologie. Théorie, méthode et applications*. Presses de l'Université d'Ottawa, Ottawa et Armand Colin, Paris, 1998. 21, 22, 24, 25, 28, 29

Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, pages 249–254, 1996. 89

Bruno Cartoni. Lexical Morphology in Machine Translation: A Feasibility Study. Dans *EACL*, pages 130–138, 2009. 110, 122

Stéphane Chaudiron. *L'Évaluation des systèmes de traitement de l'information textuelle: vers un changement de paradigme*. mémoire pour hdr, Université de Paris X, Paris, 2001. 46

Aitao Chen et Fredric Gey. Translation term weighting and combining translation resources in cross-language retrieval. Dans *Proceedings of TREC Conference*, 2001. 59

Kenneth Ward Church et Patrick Hanks. Word Associations, Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 38

Vincent Claveau et Ewa Kijak. Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé. Dans *Conférence Traitement automatique des langues naturelles TALN*, Montréal, Québec, Canada, 2010. 110

Anne Condamines. Linguistique de corpus et terminologie. *Langages*, 39(157):36–47, 2005. 22, 27

Mathias Creutz et Krista Lagus. Unsupervised Discovery of Morphemes. Dans *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, pages 21–30, 2002. 72

Fabienne Cusin-Berche. *Les mots et leurs contextes*. Presses Sorbonne Nouvelle, 2003. 62, 65

B. Daille, B. Habert, C. Jacquemin, et J. Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257, 1996. 29, 30, 123

Béatrice Daille. *Approche mixte pour l'extraction de terminologie: statistiques lexicales et filtres linguistiques*. These, Université Paris VII, 1994. 135

Béatrice Daille. Terminology Mining. Dans M.T. Pazienza, éditeur, *Information Extraction in the Web Era*, pages 29–44. Springer, Paris, 2003. 30

Béatrice Daille. Variations and application-oriented terminology engineering. *Terminology*, 11(1):181–196, 2005. 30, 31, 47

Béatrice Daille et Helena Blancafort. Knowledge-poor and knowledge-rich approaches for multilingual terminology extraction. Dans *Proceedings, 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, page 14p, Samos, Greece, 2013. 94

Estelle Delpech. *Traduction assistée par ordinateur et corpus comparables : contributions à la traduction compositionnelle*. These, Université de Nantes, Juillet 2013. URL http://hal.archives-ouvertes.fr/tel-00905930. 98, 101, 102, 111, 117, 118

Amélie Depierre. Souvent HAEMA varie… : Les dérivés du grec HAEMA en anglais : Étude de cas de variation. *Terminology*, 13(2):155–176, 2007. 29, 30

Dictionnaire Collins. Collins Dictionary, 2013. URL http://www.collinsdictionary.com/dictionary/english/work. 65

Wolfgang U. Dressler et Lavinia Merlini Barbaresi. *Morphopragmatics: Diminutives and Intensifiers in Italian, German, and Other Languages*. Walter de Gruyter, Berlin, 1994. 68

Claude Dubois. *Le petit Larousse illustré*. Larousse, Paris, 1980. 21

Chris Dyer. Using a maximum entropy model to build segmentation lattices for MT. Dans *Proceedings of HLT-NAACL 2009*, 2009. 73, 75, 78, 107

Hervé Déjean et Éric Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, 2002. URL http://lexicometrica.univ-paris3.fr/thema/thema6/Dejean.pdf. 27

Bernard Fradin. *Nouvelles approches en morphologie*. Presses Universitaires de France, Paris, 2003. 63

Alexander Fraser, Marion Weller, Aoife Cahill, et Fabienne Cap. Modeling Inflection and Word-Formation in SMT. Dans *Proceedings of EACL-12*, 2012. 107

Judit Freixa. Causes of denominative variation in terminology : a typology proposal. *Terminology*, 12(1): 51–77, 2006. 29, 30

Fabienne Fritzinger et Alexander Fraser. How to avoid burning ducks: combining linguistic analysis and corpus statistics for German compound processing. Dans *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 224–234, Stroudsburg, PA, USA, 2010. URL http://dl.acm.org/citation.cfm?id=1868850.1868884. 18, 72, 75

Oana Frunza et Diana Inkpen. Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques. Dans *International Journal of Linguistics*, 2009. 80

Joëlle Gardes-Tamine. *La Grammaire. Phonologie, morphologie, lexicologie*, volume 1. Armand Colin, Paris, 2010. 61, 67

Eric Gaussier. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. *Proceedings of COLING/ACL*, 1:444–450, 1998. 42

Natalia Grabar et Thierry Hamon. Unsupervised Method for the Acquisition of General Language Paraphrases for Medical Compounds. Dans *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 94–103, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-4812. 130

Natalia Grabar et Pierre Zweigenbaum. Lexically-based terminology structuring. *Terminology*, 10(1):25–53, 2004. 30

Gregory Grefenstette. The World Wide Web as a resource for example-based machine translation tasks. Dans *Translating and the Computer 21*, London, 1999. ASLIB. 43, 51, 109, 110

Clément De Groc. Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. Dans *The IEEEWICACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France, 2011. 139

Marie Guégan et Claude De Loupy. Knowledge-poor approach to shallow parsing: Contribution of unsupervised part-of-speech induction. Dans *Proceedings of the Conference Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, 2011. 41

Rima Harastani. *Alignement lexical en corpus comparables : le cas des composés savants et des adjectifs relationnels*. These, Université de Nantes, 2014. URL http://tel.archives-ouvertes.fr/tel-00949025. 141

Daniel Hewlett et Paul Cohen. Fully Unsupervised Word Segmentation with BVE and MDL. Dans *Proceedings of ACL 2011*, pages 540–545, Portland, Oregon, 2011. 73, 74

Susan Hunston et Geoff Thompson. *Evaluation in text: authorial stance and the construction of discourse*. Oxford University Press, New York, 2000. 25

Fidelia Ibekwe-SanJuan. A linguistic and mathematical method for mapping thematic trends from texts. Dans *Proceedings of ECAI-98*, pages 170–174, Brighton, Angleterre, 1998. 30

Ray S. Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, 1983. 25

C. Jacquemin et E. Tzoukermann. NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. Dans T. Strzalkowski, éditeur, *Natural Language Information Retrieval*, pages 25–74. Kluwer, Boston, MA., 1999. 30

Christian Jacquemin. Fastr: a unification-based front-end to automatic indexing. Dans *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO 1994)*, pages 34–47, 1994. 124

Christian Jacquemin. Syntagmatic and Paradigmatic Representations of Term Variation. Dans *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 341–348, 1999. 29, 124

Kyo Kageura et Bin Umino. Methods of automatic term recognition. *Terminology*, (3):259–289, 1996. 23, 25, 38

Adam Kilgarriff. *Polysemy*. These, University of Sussex, 1992. 39

Rostislav Kocourek. *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Oscar Brandsletter, Wiesbaden, 1991. 24

Philipp Koehn et Kevin Knight. Empirical methods for compound splitting. Dans *Proceedings of EACL 2003*, Budapest, Hungary, 2003. 18, 71, 72, 74, 75, 78, 82, 98, 100, 101, 102, 107, 111, 113

Irina Kostina. Clasificación de la variación conceptual de los términos basada en la modulación semántica discursiva [Classification of Conceptual Variation of Terms Based on the Semantic Discursive Modulation]. *Íkala*, 16(27), 2011. 29

J. R. Landis et G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174, 1977. 89

Stefan Langer. Zur Morphologie und Semantik von Nominalkomposita. Dans *Proceedings of KONVENS 1998*, pages 83–97, 1998. 68, 70, 82

Martha Larson, Daniel Willett, Joachim Köhler, et Gerhard Rigoll. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. Dans *Proceedings of the Sixth International Conference on Spoken Language Processing*, pages 945–948, Beijing, 2000. 59

Els Lefever, Lieve Macken, et Veronique Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. Dans *Proceedings of the 12th Conference of the EACL*, page 496–504, Athens, Greece, 2009. 42

Alise Lehmann et Françoise Martin-Berthet. *Introduction à la lexicologie*. Armand Colin, Paris, 2008. 28, 29, 61, 62, 63, 64, 65

Pierre Lerat. *Les langues spécialisées*. Presses Universitaires de France, Paris, 1995. 24

Marie-Claude L'Homme. *La terminologie: principes et techniques*. Paramètres. Les Presses de l'Université de Montréal, Montréal, 2004. 21, 25, 28, 39, 40, 49

Rochelle Lieber. On the lexical semantics of compounds: Non-affixal (de)verbal compounds. Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 127–145. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 83

Jacqueline Léon. Lexies, synapsies, synthèmes: le renouveau des études lexicales en France au début des années 1960. Dans Sergio Scalise et Irene Vogel, éditeurs, *History of Linguistics in Texts and Concepts*, pages 405–418. Nodus Publikationen, Münster, 2004. 62

K. Macherey, A.M. Dai, D. Talbot, A.C. Popat, et F. Och. Language-independent Compound Splitting with Morphological Operations. Dans *Proceedings of ACL 2011*, pages 1395–1404, Portland, Oregon, 2011. 59, 73, 75, 107

Bernardo Magnini et Gabriela Cavaglià. Integrating Subject Field Codes in WordNet. Dans *Proceedings of LREC 2000*, 2000. 41

Chiara Melloni et Antonietta Bisetto. Parasynthetic compounds. Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 199–217. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 65, 82

Igor Mel'čuk. *Cours de morphologie générale*, volume 4. Les Presses de l'Université de Montréal, Montréal, 1997. 68

E. Morin, B. Daille, K. Takeuchi, et K. Kageura. Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. Dans *Proceedings of ACL 2007*, pages 664–671, Prague, 2007. 40, 44, 46, 113

Emmanuel Morin et Béatrice Daille. Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*, pages 141–154, 2012. 44, 51

Widad Mustafa El Hadi, Ismail Timimi, et Marianne Dabbadie. EVALDA-CESART Project: Terminological Resources Acquisition Tools Evaluation Campaign. Dans *Proceedings of LREC 2004*, pages 515–518, Lisbon, Portugal, 2004. 45, 48

Widad Mustafa El Hadi, Ismail Timimi, Marianne Dabbadie, Khalid Choukrie, Olivier Hamon, et Yun-Chuang Chiao. Terminological Resources Acquisition Tools: Toward a User-oriented Evaluation Model. Dans *Proceedings of LREC 2006*, pages 515–518, 2006. 45, 52

Fiammetta Namer. Automatiser l'analyse morpho-sémantique non affixale: le système DériF. *Cahiers de grammaire*, 28:31–48, 2003. 70, 136

Fiammetta Namer. *Morphologie, lexique et traitement automatique des langues*. Lavoisier, Paris, 2009. 64, 84

Franck Neveu. *Dictionnaire des sciences du langage*. Armand Colin, Paris, 2004. 28, 63

Gabriel Otman. *Les représentations sémantiques en terminologie*. Masson, Paris, 1996. 21, 22, 23

Niela Ott. Evaluation of the bananasplit compound splitter, 2006. URL `http://niels.drni.de/n3files/bananasplit/Evaluation-CompoundSplitter.pdf`. 71, 74, 75

Niels Ott. Measuring semantic relatedness of german compounds using germanet, 2005. URL `http://niels.drni.de/n3files/bananasplit/Compound-GermaNet-Slides.pdf`. 70, 82

K. Papineni, S. Roukos, T. Ward, et W. Zhu. BLEU: a Methode for Automatic Evaluation of Machine Translation. Dans *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania, 2002. 135

Mojca Pecman. A study of neology as a rhetorical device in scientifique papers. *Terminology*, 18(1):27–58, 2012. 28

Vito Pirrelli, Emiliano Guevara, et Marco Baroni. Computational issues in compound processing. Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 271–285. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 108

Vladimir Plungian. *Obshchaja morfologija: Vvedenie v problematiku [General morphology: An introduction]*. Editorial URSS, Moscow, 2000. 62, 68

Bernard Pottier. Introduction à l'étude des structures grammaticales fondamentales. *La Traduction Automatique*, III-3:63–91, 1962. 62

Angela Ralli. *Compounding in Modern Greek*. Springer, 2013. 68

Reinhard Rapp. Identify word translations in non-parallel texts. Dans *Proceedings of the 35th annual meeting of the association for computational linguistics (ACL 1995)*, page 320–322, Boston, MA, USA, 1995. 43

Reinhard Rapp. Automatic Identification of Word Translation from Unrelated English and German Corpora. Dans *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL 1999)*, page 519–526, College Park, Maryland, USA, 1999. 39, 43, 51

Josette Rey-Debove. *Sémiotique*. Lexique. Presses Universitaires de France, Paris, 1979. 21

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, et Takehito Utsuro. Compiling French-Japanese Terminologies from the Web. Dans *Proceedings of the 11th conference of the European chapter of the association for computational linguistics*, EACL'06, pages 225–232, 2006. 108, 109, 110

Thomas Roeper et Muffy E. A. Siegel. A Lexical Transformation for Verbal Compounds. *Linguistic Inquiry*, 9(2):199–260, 1978. 66

Guy Rondeau. *Introduction à la terminologie*. Gaëtan Morin, Chicoutimi, 1983. 28

J.C. Sager, D. Dungworth, et P.F. Mac Donald. *English Special Languages*. Oscar Brandsletter, Wiesbaden, 1980. 24

Edward Sapir. The Status of Linguistics as a Science. Dans D. G. Mandelbaum, éditeur, *Culture, Language and Personality (1958)*. University of California Press, Berkeley, CA, 1929. 22

Sergio Scalise et Antonio Fabregas. The head in compounding. Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 109–125. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 82

Sergio Scalise et Irene Vogel. Why compounding? Dans Sergio Scalise et Irene Vogel, éditeurs, *Cross-disciplinary issues in compounding*, volume 311 de *Current issues in linguistic theory*, pages 1–18. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2010. 63, 64, 87

Helmut Schmid, Arne Fitschen, et Ulrich Heid. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. Dans *Proceedings of LREC 2004*, pages 1263–1266, Lisbon, Portugal, 2004. 70, 72, 82, 136

John McH. Sinclair. EAGLES. Preliminary Recommendations on Corpus Typology, 1996. URL `www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpustyp.ps.gz`. 27

Jonas Sjöbergh et Viggo Kann. Finding the correct interpretation of Swedish compounds, a statistical approach. Dans *Proceedings of LREC 2004*, pages 899–902, Lisbon, 2004. 70

Sara Stymne. German compounds in factored statistical machine translation. Dans *Proceedings of GoTAL – 6th International Conference on Natural Language Processing*, pages 464–475, Gothenburg, 2008. 72, 82

Sara Stymne, Nicola Cancedda, et Lars Ahrenberg. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. *Computational Linguistics*, 39(4):1067–1108, 2013. 72, 75, 107

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambrigde, 1990. 24

Takaaki Tanaka et Timothy Baldwin. Translation Selection for Japanese-English Noun-Noun Compounds. Dans *Proceedings of Machine Translation Summit IX*, pages 378–385, New Orleans, USA, 2003. 109

Fabienne Ville-Ometz, Jean Royauté, et Alain Zasadzinski. Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology*, 13(1):61–84, 2007. 124, 128

Spela Vintar. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16:141–158, 2010. 42, 44, 46, 52, 110, 122

Sami Virpioja, Oskar Kohonen, et Krista Lagus. Unsupervised Morpheme Analysis with Allomorfessor. Dans *CLEF'09*, pages 609–616, 2009. 73, 78, 100, 133

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, et Mikko Kurimo. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline, Technical Report. Dans *Aalto University publication series SCIENCE + TECHNOLOGY, 25/2013*, Aalto University, Helsinki, 2013. URL `https://aaltodoc.aalto.fi/handle/123456789/11836`. 74, 98

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, et Mikko Kurimo. Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology. *Traitement Automatique des Langues*, 52(2):45–90, 2011. 52

Jorge Vivaldi et Horacio Rodríguez. Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2):225–248, 2007. 25

Jorge Vivaldi et Horacio Rodríguez. Finding Domain Terms using Wikipedia. Dans *Proceedings of the 7th LREC*, 2010a. 41

Jorge Vivaldi et Horacio Rodríguez. Using Wikipedia for term extraction in the biomedical domain : first experience. Dans *Procesamiento del Lenguaje Natural 45*, pages 251–254, 2010b. 41

Jorge Vivaldi et Horacio Rodríguez. Using Wikipedia for Domain Terms Extraction. Dans *Proceedings of TKE'12*, 2012. 41

M. Weller et U. Heid. Analyzing and Aligning German Compound Nouns. Dans *Proceedings of LREC 2012*, Istanbul, 2012. 59, 72, 82, 107, 111, 117, 120, 122, 136

Marion Weller, Helena Blancafort, Anita Gojun, et Ulrich Heid. Terminology extraction and term variation patterns: a study of French and German data. Dans *Proceedings of German Society for Computational Linguistics and Language Technology (GSCL 2011)*, Hamburg, Germany, 2011. 41, 123, 124, 130

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, et Alexander Fraser. Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. Dans *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5709. 111, 113

Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press, 1956. 22

Eugen Wüster. *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik. (Die nationale Sprachnormung und ihre Verallgemeinerung)*. VDJ, Berlin, 1931. 22

F. Yoshikane, K. Tsuji, K. Kageura, et C. Jacquemin. Detecting Japanese Term Variation in Textual Corpus. Dans *Proceedings of 4th International Workshop on Information Retrieval with Asian Languages (IRAL 1999)*, pages 97–108, Taipei, Taiwan, 1999. 123, 124, 130

Andrej A. Zaliznjak. *Grammaticheskij Slovar' Russkogo Jazyka [Grammatical Dictionary of the Russian Language]*. Russkij jazyk, Moscow, 1977. 83

l'université nantes angers le mans
COMMUNAUTÉ D'UNIVERSITÉS ET ÉTABLISSEMENTS

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Elizaveta LOGINOVA CLOUET

Traitement automatique des termes composés : segmentation, traduction et variation

Processing of Compound Terms: Segmentation, Translation and Variation

Résumé

Le nombre de termes spécialisés croît constamment dans les documents, à un rythme difficile à suivre pour les organismes de normalisation de la terminologie. Les méthodes de construction des lexiques terminologiques bilingues à partir de corpus de textes proposent des solutions. Notre thèse s'inscrit dans cette problématique : la construction de lexiques bilingues à partir de corpus comparables.

Les termes composés (les termes contenant plusieurs radicaux, mais un seul mot graphique) constituent un défi pour les applications du traitement automatique des langues. Étant donné leur forme graphique, ils sont souvent traités comme des termes simples, ce qui empêche de capturer leur complexité sémantique. Notre participation à une évaluation d'extraction automatique de termes a permis de vérifier notre hypothèse : les termes composés nécessitent un traitement particulier dans un contexte multilingue.

Nous avons proposé une méthode de reconnaissance et de segmentation des termes composés, combinant des caractéristiques dépendantes et indépendantes de la langue. Elle permet d'obtenir des résultats comparables à ceux des méthodes de l'état de l'art, tout en étant validée sur un échantillon de familles de langues varié (germanique, slave, romane) et adaptable au domaine de spécialité (vérifiée sur deux domaines : l'énergie éolienne et le cancer du sein).

Nous avons exploité les segmentations produites pour la traduction compositionnelle des termes et pour la détection des variantes syntagmatiques des termes composés dans les textes spécialisés. Ces deux expériences illustrent l'utilité de la segmentation pour la construction des lexiques terminologiques bilingues.

Mots clés

terminologie computationnelle, extraction terminologique, lexique terminologique bilingue, termes composés, composition morphologique, segmentation des mots composés, traduction compositionnelle, variation terminologique.

Abstract

The number of specialized terms continuously grows in the documents, at a pace which is difficult to follow for the terminology standardization organizations. The methods of bilingual term lexicon construction from the text corpora provide solutions. Our thesis falls into this topic: bilingual lexicon acquisition from comparable corpora.

Compound terms (terms containing several roots, but a single graphical unit) are challenging for natural language processing applications. Given their graphical form, they are often handled in the same manner as single word terms, which prevents from apprehending their semantic complexity.

Our involvement in an automatical terminology extraction evaluation allowed us to check our hypothesis: compound terms need a particular processing in a multilingual context.

We proposed a method for compound terms recognition and splitting, which combines language-independent and language-specific features. It allowed us to obtain results comparable with those of state-of-the-art methods, while validating on a sample of languages from several families (germanic, slavic, romance languages), and adapting the method to specialized domains (tested on two domains: wind energy and breast cancer).

We used the produced segmentations for compositional translation of compound terms, and for their multi-word variant recognition in the specialized texts. These two experiments illustrate that compound splitting is beneficial for the bilingual term lexicon acquisition task.

Key Words

computational terminology, terminology extraction, bilingual term lexicon, compound terms, morphological compounding, compound splitting, compositional translation, term variation.